

MARGHERITA FRANCESCATTO

**DISSECTING HUMAN CNS GENE EXPRESSION**

Tese de Candidatura ao grau de Doutor em Biologia Básica e Aplicada do ICBAS submetida ao Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto.

Orientador – Peter Heutink

Categoria – Professor

Afiliação – Deutsches Zentrum für Neurodegenerative Erkrankungen

Coorientador – Alexandre do Carmo

Categoria – Professor

Afiliação – Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto.



“If our brains were simple enough for us to understand them,  
we’d be so simple that we couldn’t.”  
– Ian Stewart



## Table of Contents

Summary.....	7
Resumo.....	9
List of publications.....	11
Abbreviations and symbols used in the thesis.....	13
Chapter 1: General introduction.....	15
1.1 Introduction: regional vulnerability in neurodegenerative diseases.....	17
1.2 The Human Genome Project: putting the basis for genome-wide expression profiling.....	22
1.3 Understanding transcriptional regulation: ENCODE and FANTOM.....	23
1.3.1 ENCODE: ENCyclopedia Of DNA Elements.....	24
1.3.2 FANTOM: Functional ANnoTation Of Mammals.....	27
1.4 Outlook and thesis aims.....	30
1.5 Thesis outline.....	31
Chapter 2: Regional differences in gene expression and promoter usage in aged human brains.....	33
Chapter 3: A promoter level mammalian expression atlas.....	47
Chapter 4: A high resolution spatial promoterome of the human brain.....	59
Chapter 5: Brain-specific noncoding RNAs are likely to originate in repeats and may play a role in up-regulating genes in cis.....	95
Chapter 6: General discussion.....	117
6.1 Results summary.....	119
6.2 Discussion.....	121
6.3 Limitations.....	126
6.4 Future directions.....	128
6.5 Conclusions.....	130
Acknowledgments.....	131
References.....	135



## Summary

The central nervous system is the most complex organ of the human body, composed of numerous anatomical regions characterized by different cellular compositions and functions, and interconnected by complex communication networks. One of the most tragic conditions of this organ is given by neurodegenerative diseases, that affect millions of people worldwide and are an increasing burden for the modern society as the population ages. One of the most striking characteristics of most neurodegenerative diseases is that degeneration seems to affect specific regions and/or cellular populations of the central nervous system. Although theories exist, the mechanisms underlying this regional vulnerability remain largely unknown. It is likely that transcriptional networks active in specific areas of the brain or transcripts expressed in a region-specific way are involved in the process. A crucial step to verify this is to establish solid knowledge on the transcriptional features characteristic of the aged central nervous system and its districts. The general aim of this thesis was to gain insight into the dynamics of transcription in the aged central nervous system and specifically create a high resolution expression profile atlas of distinct brain regions from aged donors, to be eventually compared with material derived from patients affected by neurodegenerative diseases. In Chapter 2, we present a pilot study where we profile transcription in 5 anatomical regions of the central nervous system. In Chapter 3, as part of the FANTOM5 consortium, we participate in the creation of a large expression atlas encompassing a broad array of human and mouse primary cells, cell lines and tissues. In Chapter 4, we focus on the 15 central nervous system regions included in the FANTOM5 tissue collection, representing an expansion of the pilot presented in Chapter 2. In Chapter 5 we use a previously published custom microarray non-coding RNA expression dataset generated from twelve human tissues to identify brain-specific non-coding RNAs and investigate their characteristics. Overall, we provide evidence of specific transcriptional features that characterize the human central nervous system and identify large arrays of poorly characterized transcripts that are expressed in specific regions and might be involved in regional vulnerability in neurodegenerative diseases.





## Resumo

O sistema nervoso central é o órgão mais complexo do corpo humano, composto por várias regiões anatómicas caracterizadas por diferentes composições celulares e funções, e interligadas por redes de comunicação complexas. Uma das condições mais trágicas do sistema nervoso central é dada pelas doenças neurodegenerativas, que afetam milhões de pessoas em todo o mundo e são um fardo crescente para a sociedade moderna, por causa do envelhecimento da população. Uma das características mais marcantes da maioria das doenças neurodegenerativas é que a degeneração parece atacar regiões e/ou populações celulares específicas do sistema nervoso central. Embora haja teorias, os mecanismos subjacentes a esta vulnerabilidade local são mal compreendidos. É provável que redes de transcrição ou transcritos ativos em áreas específicas do cérebro estejam envolvidos nesta vulnerabilidade e perda neuronal localizada. Um passo decisivo para testar esta hipótese é estabelecer uma sólida compreensão das características de transcrição que são típicas do sistema nervoso central idoso e das suas partes. O objetivo geral deste trabalho foi o de obter conhecimento detalhado sobre a dinâmica de transcrição no sistema nervoso central idoso e, em particular, criar um mapa de alta resolução da expressão gênica em diferentes regiões anatómicas do sistema nervoso central, a ser comparado com dados de doadores que sofrem de doenças neurodegenerativas. No Capítulo 2, apresentamos um estudo piloto em que analisamos a expressão gênica em 5 regiões do sistema nervoso central. No Capítulo 3, como parte do consórcio FANTOM5, participamos na criação de um grande atlas da expressão gênica que inclui dados provenientes de uma variedade exaustiva de linhas celulares e tecidos humanos e murinos. No Capítulo 4, concentramo-nos nas 15 regiões do sistema nervoso central disponíveis no contexto do consórcio FANTOM5, que representam uma expansão do estudo piloto apresentado no Capítulo 2. No Capítulo 5, usamos dados publicados anteriormente sobre a expressão de RNA não-codificante em doze tecidos humanos para identificar RNAs não-codificantes específicos do cérebro e investigar as suas características. No geral, nós fornecemos evidência de características de

transcrição específicas que caracterizam o sistema nervoso central humano e identificamos grandes matrizes de transcritos mal caracterizados que são expressos em regiões específicas e podem estar envolvidos na vulnerabilidade regional em doenças neurodegenerativas.

## List of publications

Pardo LM\*, Rizzu P\*, Francescato M, Vitezic M, Leday GG, Simón-Sánchez J, Khamis A, Takahashi H, van de Berg WD, Medvedeva YA, van de Wiel MA, Daub CO, Carninci P, Heutink P. 2013. **Regional differences in gene expression and promoter usage in aged human brains.** *Neurobiol. Aging* **34**(7):1825-36.

Forrest ARR, Kawaji H, Rehli M, Baillie K, de Hoon MJL, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, Andersson R, Mungall CJ, Meehan TF, Schmeier S, Bertin N, Jørgensen M, Dimont E, Arner E, Schmidl C, Schaefer U, Medvedeva YA, Plessy C, Vitezic M, Severin J, Semple CA, Ishizu Y, Young RS, Francescato M, et al. 2014. **A promoter level mammalian expression atlas.** *Nature* **507**:462–470.

Francescato M\*, Vitezic M\*, Rizzu P, Simón-Sánchez J, Andersson R, Kawaji H, Itoh M, Kondo N, Lassmann T, Kawai J, Suzuki H, Hayashizaki Y, Daub CO, Sandelin A, de Hoon MJL, Carninci P, Forrest ARR, Heutink P and the FANTOM consortium. 2014. **A high resolution spatial promoterome of the human brain.** In preparation.

Francescato M, Vitezic M, Heutink P and Saxena A. 2014. **Brain-specific noncoding RNAs are likely to originate in repeats and may play a role in up-regulating genes in cis.** Accepted in *Int. J. Biochem. Cell. Biol.*

Vavoulis D, Francescato M, Heutink P and Gough J. 2014. **DGEclust: differential expression analysis of clustered count data.** Under review in *Genome Biol.*

\* These authors contributed equally.



## Abbreviations and symbols used in the thesis

3D = three dimensional

3C = chromosome conformation capture

5C = chromosome conformation capture carbon copy

AD = Alzheimer's Disease

ALS = amyotrophic lateral sclerosis

*APP* = amyloid-beta precursor protein, gene

bp = base pair

CAGE = cap analysis of gene expression

cDNA = complementary DNA

CGI = CpG Island

ChIA-PET = chromatin interaction analysis with paired-end-tag sequencing

ChIP = chromatin immunoprecipitation

CNS = central nervous system

DNA = deoxyribonucleic acid

ENCODE: encyclopedia of DNA elements

FAIRE = formaldehyde assisted isolation of regulatory elements

FANTOM = functional annotation of mammals

FTD = frontotemporal dementia

*FUS* = fused in sarcoma, gene

GABA = gamma-Aminobutyric acid

GO = Gene Ontology

GWAS = genome-wide association study

H3K4me1 = monomethylated histone H3 lysine 4, histone modification

H3K27ac = acetylated histone 3 lysine 27, histone modification

H3K9ac = acetylated histone 3 lysine 9, histone modification

HD = Huntington's Disease

HGP = human genome project

*HTT* = huntingtin, gene

lncRNA = long non-coding RNA

*MAPT* = microtubule-associated protein tau, gene

Mb = megabase

ncRNA = non-coding RNA

PD = Parkinson's Disease

RIN = RNA integrity number

RNA = ribonucleic acid

*SNCA* = alpha-synuclein, gene

SNP = single nucleotide polymorphism

*SOD1* = superoxide dismutase 1, gene

*TARDBP* = TAR DNA Binding Protein, gene, also known as *TDP-43*

TF = transcription factor

TSS = transcription start site

# **Chapter 1**

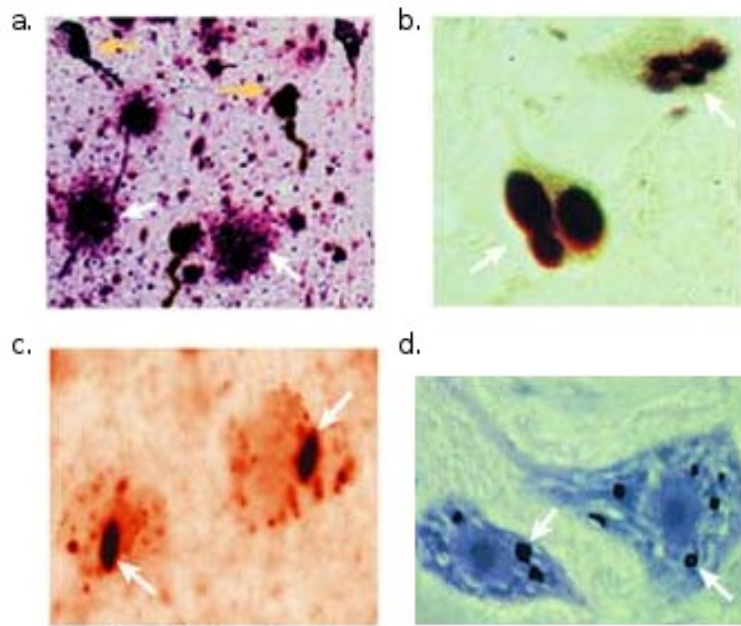
## **General introduction**





## 1.1 Introduction: regional vulnerability in neurodegenerative diseases

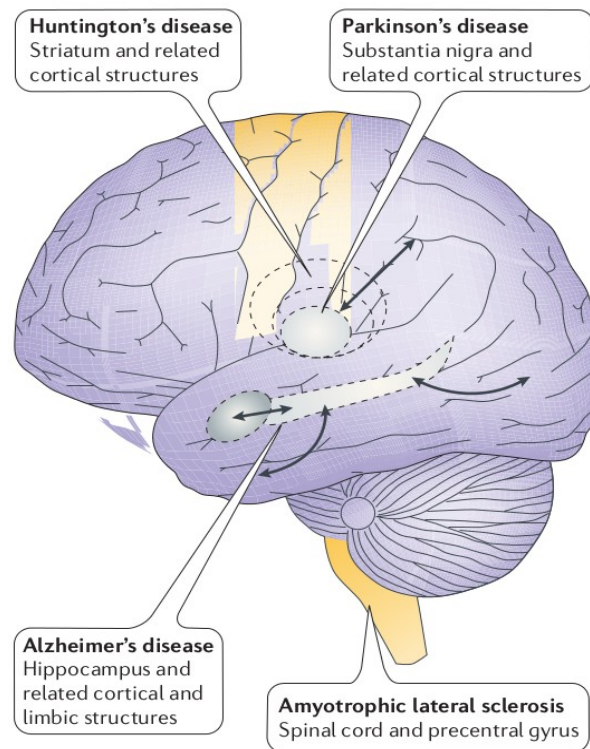
Neurodegenerative diseases represent a large group of hereditary and sporadic neurological disorders with heterogeneous clinical and pathological manifestations, characterized by advancing central nervous system (CNS) dysfunction associated to the progressive decay and eventually death of neurons (Przedborski et al. 2003). They include diseases such as Alzheimer's Disease (AD), Frontotemporal Dementia (FTD) and other dementias, Parkinson's Disease (PD), Amyotrophic Lateral Sclerosis (ALS) and Huntington's Disease (HD). Although heterogeneous in their clinical presentation, age of onset, duration and progression, certain pathways and biological processes appear to be consistently altered. Studies in animal models and patient post-mortem material provide strong evidence of increased oxidative stress and impaired mitochondrial function (Lin and Beal 2006; Johri and Beal 2012), axonal transport defects (Millecamps and Julien 2013), defects in the autophagy (Nixon 2013) and mitophagy (Palikaras and Tavernarakis 2012; Ashrafi and Schwarz 2013) pathways, endoplasmic reticulum stress and unfolded protein response (Matus et al. 2011; Hetz and Mollereau 2014). It is also well accepted that CNS inflammation has a role in the progression of neurodegenerative diseases and although it may not typically represent an initiating factor, there is emerging evidence that sustained inflammatory responses involving microglia and astrocytes contribute to disease progression (Glass et al. 2010; Cunningham 2013). Finally the recent discovery that mutations in the RNA-binding proteins *TARDBP* and *FUS* are causal for up to 8% of the familial cases of ALS (Lagier-Tourenne 2010) suggests that improper RNA processing might be involved in the pathogenesis and progression of at least certain neurodegenerative diseases, such as ALS and FTD. This hypothesis is further supported by the observation that a pathological feature of ALS and FTD patients carrying the *c9orf72* repeat expansion mutation show sequestration of RNA binding proteins in RNA foci present in the nucleus or cytoplasm of cells (Lagier-Tourenne et al. 2013).



**Figure 1. Examples of protein aggregates identified in distinct neurodegenerative diseases.** Amyloid plaques (white arrows) and neurofibrillary tangles (yellow arrows) are typically found in post-mortem brain material of AD patients (a). PD and ALS are generally characterized by cytoplasmic aggregates (b. and d.) while in HD intranuclear aggregates are found. Adapted from (Soto 2003).

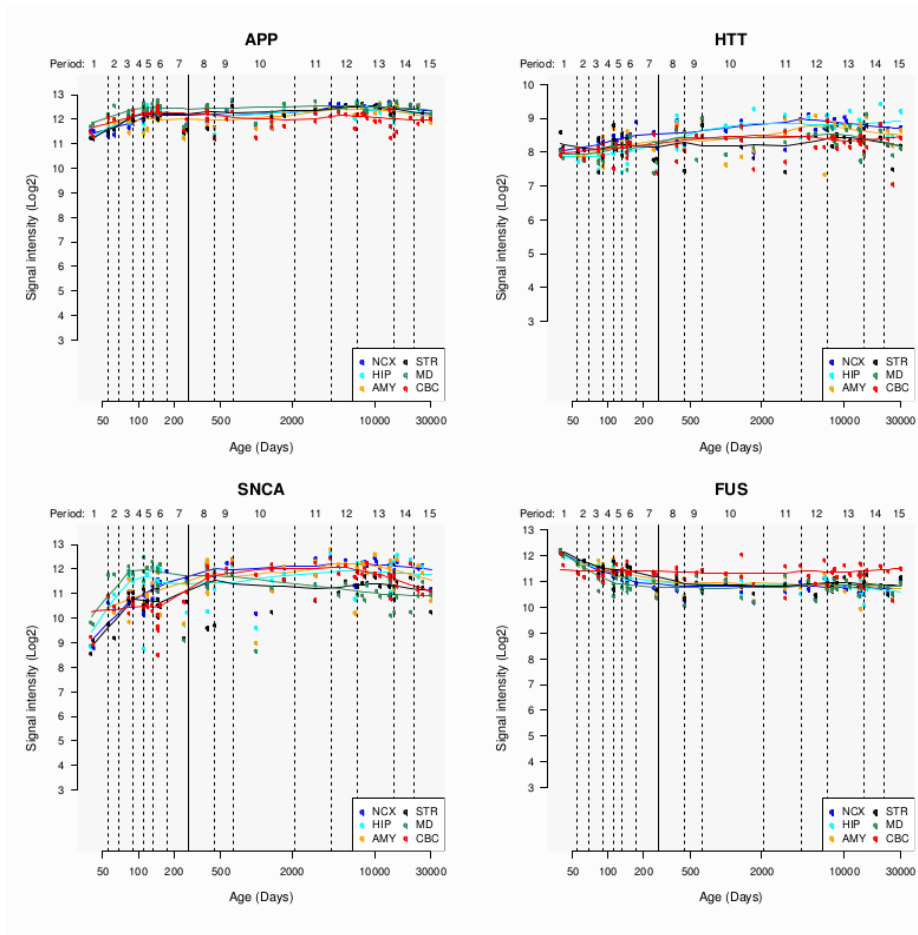
Even though the clinical manifestation of all these diseases is diverse, at the molecular level they often share the phenomenon of accumulation of abnormally folded proteins in the form of small oligomers, aggregates or large-protein inclusions. The accumulation of distinct protein-based macroscopic deposits is therefore a hallmark of neurodegenerative diseases and the composition and distribution of the deposits is a defining property of each of them (Figure 1). For example AD is characterized by extracellular amyloid plaques and intracellular neurofibrillary tangles (Gamblin et al. 2003); PD by characteristic intraneuronal cytoplasmic inclusions – termed Lewy Bodies – composed of several proteins (Dauer and Przedborski 2002); distinct subtypes of FTD by tau- or ubiquitin-positive deposits in neurons and glia (Bigio 2013); HD by intracellular aggregates called inclusion bodies (IBs). Some of the mutations associated with familial cases of these diseases affect the major protein components of the deposits: e.g. the

identification of mutations in the alpha-synuclein gene (*SNCA*) in familial forms of PD (Polymeropoulos et al. 1997) eventually led to the discovery of *SNCA* as principal component of Lewy bodies. Analogously the beta-amyloid precursor protein (*APP*), first causal gene identified for AD (Goate et al. 1991), is a major component of amyloid plaques characteristic of AD (Glenner and Wong 1984). Additionally FTD familial cases associated to *MAPT* mutations show neuronal and glial deposits staining positive for *MAPT* (Hutton et al. 1998) while in familial forms of ALS associated with *FUS* mutations abnormal cytoplasmic glial and neuronal inclusions staining positive for *FUS* were observed (Lagier-Tourenne et al. 2010). Finally the IBs that characterize HD stain positive for mutant huntingtin (*HTT*) and ubiquitin (Arrasate and Finkbeiner 2012). In light of these and similar observations, neurodegenerative diseases are currently viewed as cerebral proteopathies, in which the accumulation of particular proteins is a key causative factor (Haass and Selkoe 2007). Another common feature of most neurodegenerative diseases is that deposit formation, pathology and eventually neuronal loss is restricted to a limited number of brain regions or subsets of neurons (Saxena and Caroni 2011; Jackson 2014) (Figure 2). In AD initial symptoms are related to prominent memory impairment and this correlates to focused neurodegeneration in hippocampus and parahippocampal gyrus (Hyman 1984). PD predominantly manifests clinically as a movement disorder and is associated to the initial degeneration and loss of dopaminergic neurons in the substantia nigra (Sulzer and Surmeier 2013). HD manifests as well as a movement disorder, however neuronal loss is initially restricted to the GABAergic neurons of striatum (Ross and Tabrizi 2011). ALS is characterized by the selective and progressive loss of upper and lower motor neurons of the brainstem, spinal cord and cerebral cortex (Robberecht and Philips 2013). The mechanisms underlying this selective vulnerability remain largely unknown: one immediate explanation would be that the genes specifically involved in protein aggregation in distinct neurodegenerative diseases are expressed at higher levels in the areas that are affected the most. However, this hypothesis is easily challenged by the observation that e.g. *APP*, *HTT*, *SNCA* and *FUS* have essentially similar levels of expression in both affected and unaffected areas (Figure 3).



**Figure 2. Schematic representation of regional vulnerability in neurodegenerative diseases.** In AD neurodegeneration initially affects the hippocampus and parahippocampal gyrus. In PD initial degeneration and neuronal loss is localized to the substantia nigra. In HD neuronal loss is initially restricted to the GABAergic neurons of striatum. ALS is characterized by the selective and progressive loss of upper and lower motor neurons of the brainstem, spinal cord and cerebral cortex. Adapted from (Mattson and Magnus 2006).

An intriguing alternative explanation is that other genes directly or indirectly interacting with the ones that are mutated are differently expressed in the most vulnerable regions and contribute to a localized alteration of either expression or physical/functional properties of the mutated genes. This suggests that to achieve a broader understanding of regional vulnerability in neurodegenerative diseases, a crucial step is to move from single genes to genome-wide expression profiling approaches, that can survey simultaneously all genes expressed in one sample and find differences in expression in an unbiased way, shedding light into region specific regulation of transcription by identifying networks of co-regulated genes.



**Figure 3. Expression patterns across distinct regions of the human CNS and across different ages for four genes linked to familial forms of neurodegenerative diseases.** For each panel, the x-axis represents age (days) and the y-axis represents normalized expression (as Log2 of signal intensity). Each dot represents a sample, each color represents a region of the CNS and solid lines summarize the expression profiles for the genes APP, HTT, SNCA and FUS. The expression of these genes is essentially similar across regions affected and non-affected by neurodegeneration in the corresponding disease. Data from Human Brain Transcriptome (<http://hbatlas.org/>).

## **1.2 The Human Genome Project: putting the basis for genome-wide expression profiling**

One of the major scientific achievements of the last century, fundamental to put the basis for genome-wide expression profiling studies, was the sequencing of the full human genome. Under the name of Human Genome Project (HGP) several groups from several countries in the world joined a collaborative public effort (conceived in 1984 and officially started in 1990) with the primary aim of determining the nucleotide sequence of the entire human nuclear genome and discovering all human genes. The public effort was paralleled in 1998 by a private company (Celera Genomics of Maryland, USA), aiming at reaching the same result with a faster and more cost-effective approach (Brown 2002). Both projects concluded successfully with the release of human genome working drafts in 2001 (Lander et al. 2001; Venter et al. 2001), completed in the definitive version in 2004 (International Human Genome Sequencing Consortium 2004). The accomplishment of the HGP started the field of genomics and dramatically contributed to shaping several fields of biology into the form they have now, from basic biology, to comparative and medical genomics. In basic biology, it reshaped our view of the genome physiology, including a precise definition of the number, distribution and structure of protein-coding genes, the discovery of novel classes of non-coding RNAs (ncRNAs) and the completely unexpected pervasiveness of transposon-derived sequences, accounting for up to 45% of the genomic sequence (Lander et al. 2001). Along with the closely following sequencing of genomes of other species (such as mouse (Mouse Genome Sequencing Consortium 2002), rat (Gibbs et al. 2004) and chimpanzee (Chimpanzee Sequencing and Analysis Consortium 2005)) it boosted the field of comparative genomics, which brought e.g. to the surprising discovery that while the exomes of human and mouse are extremely similar, a substantial excess of conserved sequence, likely functional, does not code for proteins (Mouse Genome Sequencing Consortium 2002). The accomplishment of the HGP also greatly pushed forward disease research: when the project was launched, less than 100

Mendelian disease genes had been identified. With the genetic and physical maps created in the first stages of the HGP the list quickly began to grow and a decade after more than 2,850 Mendelian disease genes had been identified (Lander 2011). Similar advances were seen in uncovering the basis of common diseases: as of 2000, only about a dozen genetic variants (outside the HLA locus) had been reproducibly associated with common disorders; a decade later, more than 1,100 loci affecting more than 165 diseases and traits had been associated with common traits and diseases (Lander 2011; Naidoo et al. 2011). Overall the greatest impact of genomics has been the ability to investigate biological phenomena in a comprehensive, unbiased, hypothesis-free manner, also thanks to the creation since of several publicly accessible databases collecting information about genes (e.g. Ensembl (Flicek et al. 2014) and GENCODE (Harrow et al. 2012)), SNPs and human variation (e.g. dbSNP (Sherry 2001), HapMap (International HapMap Consortium 2003), 1000 Genomes Project database (1000 Genomes Project Consortium 2012)) and many others. As a collateral consequence, the HGP challenge directly influenced and accelerated the evolution of sequencing technology, which went paired with the decrease in sequencing costs that we are still observing. The advent of high-throughput sequencing revolutionized and became an integral part of many areas of biological research. In particular it started a new age for the study of transcriptomes and transcriptional regulation: the two major consortia working in the field and their results are outlined in the next section.

### **1.3 Understanding transcriptional regulation: ENCODE and FANTOM**

Although all cells in the human body share essentially the same genetic code, they vary hugely in their structures and functions. Sequencing the whole genome alone does not explain how this large variety is achieved starting from the same material – the DNA present in the nucleus of each cell of an individual. In the last decade

considerable efforts were made to investigate this. In particular I will outline here the main achievements of two large international consortia dedicated to the investigation of transcription and its regulation: FANTOM and ENCODE. Besides rewriting considerable chapters of schoolbook biology and giving immense insight into the biology of transcription, starting from redesigning the concept of “gene”, they produced in the years wealths of data freely accessible to the scientific community and probably daily used in many laboratories in the world, to generate and test hypotheses and design experiments.

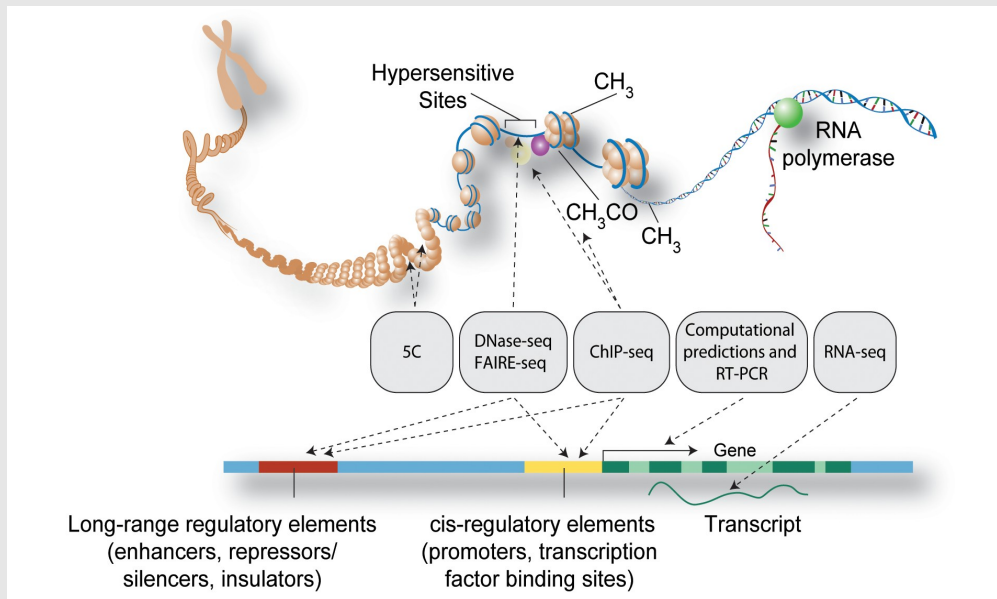
### **1.3.1 ENCODE: ENCyclopedia Of DNA Elements**

The ENCODE project was launched in September 2003 as follow-up of the HGP, with the aim of identifying all functional elements in the human genome. The ENCODE project developed in two distinct parts: a pilot (2003-2007), focusing on about 1% of the sequenced genome to test and compare existing methods to rigorously analyze a given region of the genome (ENCODE Project Consortium, Birney et al. 2007), and a first production phase (2007-2012), that scaled up the methods tested to the full genome, also thanks to the introduction in the meanwhile of next generation sequencing (ENCODE Project Consortium, Bernstein BE et al. 2012). Overall the ENCODE project efforts resulted in the generation of thousands of genome-scale data sets encompassing transcripts, sites of transcription factor (TF) binding for large arrays of TFs, DNase I hypersensitive sites, histone modifications and other functional features (Stamatoyannopoulos 2012) (the major assays at the basis of ENCODE Project are summarized in Box 1). By studying the distribution of these biochemical signatures across distinct cell types, the ENCODE projects gave immense insight into the mechanisms involved in cell-specific regulation of transcription (Arvey et al. 2012; Djebali et al. 2012; Thurman et al. 2012), the combinatorial patterns of TFs needed to achieve this precise regulation (Gerstein et al. 2012; Wang et al. 2012), and their likely genic targets (Sanyal et al. 2012; Thurman et al. 2012).



Starting from the basics, results from ENCODE and FANTOM (see also next section) redefined the unit of transcription. Although the "gene" was conventionally viewed as the fundamental unit of genomic organization, on the basis of ENCODE data it is now evident that the fundamental unit is rather the "transcript" (Washietl et al. 2007; Djebali et al. 2012). Genes represent a higher-order organizational level, in which individual transcripts are used in different cellular states, guided by differential utilization of regulatory DNA. The majority of regulatory DNA regions are highly cell-type and cell-state specific (ENCODE Project Consortium 2012; Thurman et al. 2012): considering a single cell type up to 1-2% of the DNA has regulatory function; however the frequency of regulatory DNA along the genome grows as the number of cell types and states assayed increases: it is expected that 40% and possibly more of the genome sequence encodes regulatory information (ENCODE Project Consortium 2012). The large variety of datasets produced by ENCODE led to the establishment that one of the fundamental aspects of transcriptional regulation lies in the dynamic interplay between chromatin and transcriptional machinery (Stamatoyannopoulos 2012): e.g. transcription originating from enhancer elements is predominantly detected at distal DNase I hypersensitive sites flanked by H3K4me1, H3K27ac, and H3K9ac histone modifications, as extensively documented in (Djebali et al. 2012). Additionally, the use of assays able to determine long-range chromatin interactions such as Chromosome Conformation Capture Carbon Copy (5C) (Dostie 2006) or Chromatin Interaction Analysis with Paired-End-Tag sequencing (ChIA-PET) (Fullwood et al. 2009) showed that specific physical interactions and 3D connectivity of genes with one another and with their respective controlling elements appear to be general properties of long-range regulatory control (Li et al. 2012; Sanyal et al. 2012). Finally, it is now apparent that a significant proportion of strongly disease- or trait-associated variants emerged from genome-wide association studies (GWAS) localize within regulatory DNA marked by DNase I hypersensitive sites and selected TFs (ENCODE Project Consortium 2012; Maurano et al. 2012; Schaub et al. 2012).

### Box 1: Major techniques used by the ENCODE Project.



*Graphical summary of the major techniques used by the ENCODE Project. Adapted from (ENCODE Project Consortium 2012).*

**RNA-seq:** RNA isolation, typically performed in the ENCODE Project with multiple purification protocols to separate distinct sub-cellular fractions and transcript types, followed by high-throughput sequencing.

**ChIP-seq:** Chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing. Specific regions of cross-linked chromatin, i.e. genomic DNA in complex with its bound proteins, are selected by using an antibody to a specific epitope. The enriched sample is then sequenced to determine the regions in the genome most often bound by the protein to which the antibody was directed. Most commonly used are antibodies to any chromatin-associated epitope, including transcription factors, chromatin binding proteins and specific chemical modifications on histone proteins.

**DNase-seq:** Adaptation of DNase footprinting assay to high-throughput sequencing. The DNase I enzyme preferentially cuts chromatin preparations at sites nearby bound proteins. The resulting cut points are sequenced to determine those genomic regions that are 'hypersensitive' to DNase I, corresponding to accessible DNA (also termed "open chromatin").

**FAIRE-seq:** Formaldehyde assisted isolation of regulatory elements (FAIRE). FAIRE isolates nucleosome-depleted genomic regions by exploiting the difference in crosslinking efficiency between nucleosomes (high) and sequence-specific regulatory factors (low). FAIRE consists of cross-linking, phenol extraction, and sequencing the DNA fragments in the aqueous phase.

**3C and 5C:** Chromosome Conformation Capture (3C) uses formaldehyde cross-linking to covalently trap interacting chromatin segments throughout the genome. Interacting elements are then restriction-enzyme-digested and intramolecularly ligated and the frequency with which two restriction fragments become ligated is a measure of the frequency of their interaction in the nucleus. 3C uses PCR to detect individual chromatin interactions, which is not applicable for large-scale identification of chromatin interactions. To overcome this problem, 3C-Carbon Copy (5C) uses highly multiplexed ligation-mediated amplification to first copy and then amplify parts of the 3C library, followed by detection on microarrays or by quantitative DNA sequencing.

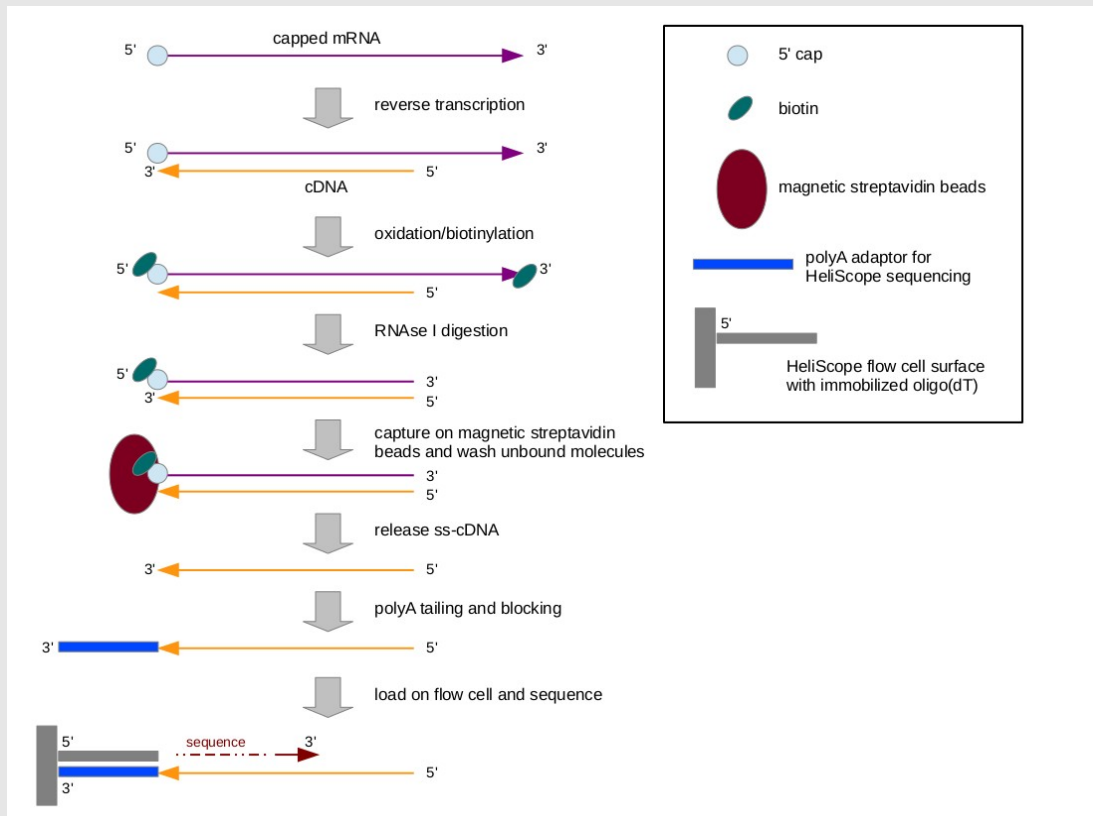
**ChIA-PET:** Chromatin Interaction Analysis with Paired-End-Tag sequencing. Combines chromatin immunoprecipitation and chromosome conformation capture to identify long-range interactions mediated by a protein of interest genome-wide, using paired-end tag libraries.

### **1.3.2 FANTOM: Functional ANnoTation Of Mammals**

To some extent parallel to the ENCODE project, FANTOM is an international research consortium established in 2000 to assign functional annotations to the full-length cDNAs that were collected during the Mouse Encyclopedia Project, established in 1995 at RIKEN (Japan), with the aim of sequencing all expressed RNAs. FANTOM has gradually developed and expanded over time to encompass the field of transcriptome analysis. The object of the project progressed from an understanding of the “elements” – the transcripts – to an understanding of the “system” – the transcriptional regulatory networks – active in individual life forms and specific to each cell. FANTOM1 and FANTOM2 projects focused on the determination of the sequences and functional annotation of large sets of full-length mouse cDNAs. The consortium cloned and annotated over 80,000 full-length cDNAs from a wide range of mouse tissues and integrated them with existing annotations, allowing for the identification of ca. 16,000 novel transcriptional units. This fundamentally contributed to the estimate of the number of genes that are part of the mouse genome, supported the innovative concept that most genes are associated to a large variety of transcripts and that alternative TSS usage and splicing are widespread phenomena (Kawai et al. 2001; Okazaki et al. 2002). Additionally the consortium reported the transcription of about 15,000 ncRNAs, of which only about 100 had been previously identified, suggesting for the first time the transcription of RNAs that do not code for proteins is a pervasive feature in mammalian genomes (Okazaki et al. 2002). In FANTOM3, besides working with full-length cDNAs, the FANTOM consortium utilized a new technology, Cap Analysis of Gene Expression (CAGE (Kodzius R et al. 2006); see also Box 2), to reveal that more than 63% of the mouse and human genomes is transcribed as RNA, instead of the ca. 1.5% fraction of protein-coding exons known at that time (Carninci et al. 2005). Additionally the expression of over 23,000 ncRNAs was confirmed and it was shown for the first time that over 73% of the transcriptional units show sense-antisense transcription (Carninci et al. 2005; Katayama et al. 2005). Work from the consortium brought to the discovery that mammalian promoters can be separated into two classes: "sharp" and "broad".

The first class represents classical promoters for which transcription initiates at a well defined position restricted to few bp; these are enriched for conserved TATA-box motifs, are usually tissue-specific and surprisingly represent a minority of the transcription start sites (TSSs). The second class, much larger, is characterized by TSSs spread across several bp; these are typically located in CpG islands and correspond to genes of broader use. Different tissues and families of genes differentially use distinct types of promoters and the usage of alternative start sites is common (Carninci et al. 2006). In FANTOM4 the focus moved to understanding how these components work together in the context of a biological network. Using CAGE adapted to high-throughput sequencing technology, the dynamics of TSS usage during a time course of monocytic differentiation in the acute myeloid leukemia cell line THP-1 was monitored. This allowed for the precise identification of active promoters and their expression levels. Computational methods were then used to build a network model of gene expression that identified the key transcriptional regulators in the differentiation process, their time-dependent activities and their target genes, which were confirmed by systematic siRNA knock-down experiments. This study was pioneering in the discovery that cell-state and cell-fate are determined by complex networks involving both positive and negative regulatory interactions among substantial numbers of TFs (FANTOM consortium et al. 2009). The latest FANTOM project, presented in Chapter 3 of this thesis, ambitiously aimed at expanding the horizon and create a map of the vast majority of human promoters and identify the regulatory networks that define virtually every single cell type in a human being.

## Box 2: CAGE – Cap Analysis of Gene Expression



Graphical representation of the CAGE protocol adapted for use with HeliScope single-molecule sequencer. Adapted from (Kanamori-Katayama et al. 2011).

Cap Analysis of Gene Expression (CAGE) is a technology developed at RIKEN, Japan, that produces a quantitative snapshot of the 5' ends of capped mRNAs in a biological sample. The ultimate output of a CAGE experiment is a set of short nucleotide sequences, often called tags, with their observed counts. The original CAGE library preparation protocol involved cDNA synthesis, cap-trapping of 5' complete cDNA/capped RNA hybrids, second-strand synthesis, linker ligation, full-length cDNA cloning in bacteria, digestion of 5' tags, and concatenation and subcloning of concatemers prior to capillary sequencing (Shiraki et al. 2003). An adaptation of the protocol for the 454 sequencer was later developed and used as leading technology for the accomplishment of the FANTOM4 project (FANTOM Consortium et al. 2009). Both the original and the 454-adapted protocols required several handling steps and PCR cycles, potentially introducing artifacts and PCR biases. The simplified HeliScope CAGE protocol, schematically represented above, aimed at reducing handling steps and avoiding PCR amplifications, to overall improve the quantitative features of the technique. Briefly, HeliScope CAGE library preparation can be summarized as follows: (a) first-strand cDNA is generated from total RNA using an excess of random primer (b) the 5' end complete first-strand cDNAs are captured through the cap structure (c) first-strand cDNA is poly(A)-tailed and blocked, then loaded directly onto the HeliScope flow cell for sequencing. An optimized protocol for CAGE library production directly applicable to Illumina sequencers was later published and is the procedure currently used in our lab (Takahashi et al. 2012).

## 1.4 Outlook and thesis aims

Work from large consortia such as ENCODE and FANTOM put the basis for the comprehensive understanding of the biology of transcription. At the same time one of the fundamental observations that emerges from these studies is that transcription is tightly regulated in a cell- and tissue-dependent manner. Considering that neurodegenerative diseases show a high degree of regional vulnerability and neuronal loss, transcriptional features, perturbations or transcripts that are specific for the regions involved may have a role in their pathogenesis. To gain a more precise understanding of region-specific regulation of transcription in brain and consequently region-specific vulnerability in neurodegenerative diseases, genome wide transcription profiling studies specifically focusing on appropriate sets of control and disease brain samples are needed. As for 2009, when this PhD project started, the only large expression profiling work performed on a comprehensive selection of different regions of the human post-mortem CNS and non-CNS tissues was (Roth et al. 2006). The main conclusions of this study were that CNS regions are significantly different from non-CNS tissues and similar between them, while the 20 CNS regions profiled could be segregated into discrete groups with underlying similarities in anatomical structure and functional activity. Besides this study, a repertoire of other expression profiling works on arrays of tissues that included brain samples suggested over the years some general features of brain transcription. In particular it was suggested that brain tissue is characterized by the highest number of genes expressed (Ramsköld et al. 2009) and by the highest transcriptional complexity (Jongeneel et al. 2005; Ramsköld et al. 2009), by over-representation of expressed simple and low-complexity repeats with respect to other tissues (Faulkner et al. 2009) and transcripts originating in CG rich regions (Roider et al. 2009). Additionally, work in mouse suggested that brain tissues express a large array of ncRNAs (Mercer et al. 2008).

With this perspective in mind, the work presented in this thesis is the first step of an ongoing effort to precisely characterize brain-specific and particularly region-

specific transcription of coding and non-coding genes in the human aged brain. This is achieved by creating a high resolution expression profiling atlas of different areas of the human aged brain, with the long term aim of investigating the networks that are transcriptionally altered in disease and functionally characterize the transcriptional networks involved.

The specific aims of this thesis can be summarized as:

- gain general insight into the dynamics of transcription in the CNS
- specifically create a high resolution expression profile atlas of distinct brain regions from aged donors
- extend the atlas to matched disease samples to identify and functionally validate networks of co-expressed transcripts perturbed in disease

## **1.5 Thesis outline**

In Chapter 2, we used CAGE to profile transcription in 5 regions of the CNS (caudate, putamen, frontal and temporal cortices, and hippocampus) derived from post-mortem material of human aged donors and additionally investigated the methylation landscape in the same regions. We first characterized the transcriptome of aged human brain and evaluated the extent of alternative promoter usage. Then, we quantified differences in gene expression and promoter usage across the 5 brain regions. Finally, we analyzed the extent to which methylation influenced the observed expression profiles.

In Chapter 3, as part of the FANTOM5 consortium, we used CAGE adapted to single molecule sequencing to map TSSs and their usage in human and mouse primary cells, cell lines and tissues to produce a comprehensive overview of gene expression across the human body.

In Chapter 4, we focused in particular on the CNS samples included in the FANTOM5 tissue collection, representing 15 regions of the human CNS, derived from post-mortem material from aged donors. First we compared the CNS expression signature and transcriptional complexity to the other tissues present in

the collection and characterized the transcriptional context of transcripts up-regulated in brain. Additionally we investigated differential expression across distinct CNS regions.

In Chapter 5 we used a previously published custom microarray ncRNA expression dataset generated from twelve human tissues to identify tissue-specific ncRNAs. We investigated the relative abundance of ncRNAs across tissues and correlated brain-specific ncRNAs expression to neighboring protein-coding genes. Additionally we investigated repeat representation at the origin and in the transcript body of brain-specific ncRNAs.

In Chapter 6 I will summarize the results presented in this thesis and discuss the major discoveries and limitations in the context of the advances in CNS-centered expression profiling studies since 2009. Additionally I'll present future applications and approaches that, making use of the work presented in this thesis, can provide further insight into brain-specific and region-specific transcriptional regulation and ultimately into the regional vulnerability that characterizes many neurodegenerative diseases.



# Chapter 2

## Regional differences in gene expression and promoter usage in aged human brains

Published as Pardo LM\*, Rizzu P\*, Francescato M, Vitezic M, Leday GG, Sanchez JS, Khamis A, Takahashi H, van de Berg WD, Medvedeva YA, van de Wiel MA, Daub CO, Carninci P, Heutink P. 2013. Regional differences in gene expression and promoter usage in aged human brains. *Neurobiol. Aging* 34(7):1825-36. \*Authors contributed equally.





## Regional differences in gene expression and promoter usage in aged human brains

Luba M. Pardo<sup>a,1</sup>, Patrizia Rizzu<sup>a,1</sup>, Margherita Francescato<sup>a,b</sup>, Morana Vitezic<sup>c,d</sup>, Gwenaël G.R. Leday<sup>e</sup>, Javier Simon Sanchez<sup>a</sup>, Abdullah Khamis<sup>f</sup>, Hazuki Takahashi<sup>c</sup>, Wilma D.J. van de Berg<sup>g</sup>, Yulia A. Medvedeva<sup>f</sup>, Mark A. van de Wiel<sup>h</sup>, Carsten O. Daub<sup>c</sup>, Piero Carninci<sup>c</sup>, Peter Heutink<sup>a,c,\*</sup>

<sup>a</sup> Section Medical Genomics, Department of Clinical Genetics, VU University Medical Center, Amsterdam, The Netherlands

<sup>b</sup> GABBA Program, Instituto de Ciências Biomédicas Abel Salazar, UP, Porto, Portugal

<sup>c</sup> RIKEN Omics Science Center, RIKEN Yokohama Institute, Yokohama, Japan

<sup>d</sup> Department of Cell and Molecular Biology (CMB), Karolinska Institute, Stockholm, Sweden

<sup>e</sup> Department of Mathematics, VU University, Amsterdam, The Netherlands

<sup>f</sup> Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

<sup>g</sup> Department of Anatomy and Neurosciences, Section Functional Neuroanatomy, Neuroscience Campus Amsterdam, VU University Medical Center, Amsterdam, The Netherlands

<sup>h</sup> Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands

## ARTICLE INFO

## Article history:

Received 27 August 2012

Received in revised form 29 November 2012

Accepted 7 January 2013

Available online 19 February 2013

## Keywords:

CAGE

Brain transcriptome

Aging

## ABSTRACT

To characterize the promoterome of caudate and putamen regions (striatum), frontal and temporal cortices, and hippocampi from aged human brains, we used high-throughput cap analysis of gene expression to profile the transcription start sites and to quantify the differences in gene expression across the 5 brain regions. We also analyzed the extent to which methylation influenced the observed expression profiles. We sequenced more than 71 million cap analysis of gene expression tags corresponding to 70,202 promoter regions and 16,888 genes. More than 7000 transcripts were differentially expressed, mainly because of differential alternative promoter usage. Unexpectedly, 7% of differentially expressed genes were neurodevelopmental transcription factors. Functional pathway analysis on the differentially expressed genes revealed an overrepresentation of several signaling pathways (e.g., fibroblast growth factor and *wnt* signaling) in hippocampus and striatum. We also found that although 73% of methylation signals mapped within genes, the influence of methylation on the expression profile was small. Our study underscores alternative promoter usage as an important mechanism for determining the regional differences in gene expression at old age.

© 2013 Elsevier Inc. All rights reserved.

## 1. Introduction

The brain is the most complex organ of the human body, and this complexity is a major landmark of human evolution (Konopka and Geschwind, 2010). The brain can be divided into different functional and anatomic regions that are established during development and maintained throughout life. The mechanisms that regulate normal brain function and differentiation are controlled by both genetic (Johnson et al., 2009) and epigenetic factors (Miller and Sweatt, 2007), and alterations in these mechanisms can lead to neurodegenerative diseases (Abdolmaleky et al., 2005). There have been tremendous advances in our understanding of the

molecular mechanisms involved in brain function, and the regional differences in these functions are beginning to be understood (Khaitovich et al., 2004; Roth et al., 2006). Less is known about the genetic mechanisms that are responsible for establishing and maintaining these differences throughout development, adulthood, and aging. Insights into these mechanisms are required to understand the differential susceptibility of distinct brain regions to neuronal insults (Double et al., 2010). For example, the genes for which mutations have been characterized in Alzheimer's disease (AD) (Joachim et al., 1989; Shen et al., 1997) and Parkinson's disease (PD) (Bandopadhyay et al., 2004) are often ubiquitously expressed whereas the observed pathology is restricted to specific brain regions and specific cell types (Double et al., 2010). Dissection of the molecular basis of this selective vulnerability will be pivotal to our understanding of disease pathogenesis and the development of specific therapies.

Much of our current insight into the molecular basis of brain function results from detailed studies of single genes or

\* Corresponding author at: Department of Clinical Genetics, VU University Medical Center, van der Boerhorststraat, 71081 BT Amsterdam, The Netherlands. Tel.: +31-205989962; fax: +31-2059983596.

E-mail address: [p.heutink@vumc.nl](mailto:p.heutink@vumc.nl) (P. Heutink).

<sup>1</sup> These authors contributed equally to this work.

molecular mechanisms often initiated by the identification of genetic mutations (Hardy and Selkoe, 2002). However, unbiased approaches, where large numbers of genes are assessed simultaneously, are expected to be more powerful to dissect the genetic mechanisms controlling brain function. Large-scale analysis of gene expression in brain was pioneered by microarray experiments (Khaitovich et al., 2004). In recent years, high-throughput sequence-based technologies have been developed to analyze the mammalian transcriptome in more detail and at greater depth (Sandelin et al., 2007). These technologies have been decisive to uncover a complex picture of the mammalian transcriptome (Carninci et al., 2005) and to identify new mechanisms of gene regulation and control of gene expression in brain (Kang et al., 2011; Tollervey et al., 2011). Among sequence-based technologies, tag-based approaches such as cap analysis of gene expression (CAGE) have been used to comprehensively profile the transcription start sites (TSSs) and the promoter regions (Takahashi et al., 2012). CAGE is a cap-trapping–based method that profiles 5' capped transcripts of both coding and noncoding RNA classes and has been pivotal in the discovery of alternatively regulated TSSs and novel regulatory elements (Caminci et al., 2006; Valen et al., 2009).

To understand how different promoters and control elements of genes establish and maintain region-specific expression patterns, we used CAGE in combination with massive parallel sequencing to profile TSSs of brain regions in 7 aged healthy individuals, at a genome-wide scale. We selected 5 samples of caudate nuclei, putamen, frontal and temporal cortices, and hippocampus, which are specifically vulnerable in the most prevalent neurodegenerative disorders (Double et al., 1996). First, we characterized the transcriptome of aged human brain and evaluated the extent of alternative promoter usage. Second, we quantified differences in gene expression and promoter usage across 5 brain regions. Finally, we analyzed the extent to which methylation influenced the observed expression profiles.

## 2. Methods

### 2.1. Brain specimens

The postmortem brain tissues were obtained from the Netherlands Brain Bank (Amsterdam, The Netherlands). The donors were aged subjects (age range: 70–91 years) without clinical signs of neurodegenerative or psychiatric disorders. All brains were neuropathologically evaluated by an experienced neuropathologist and classified for neurofibrillary tangles stage 0–VI (Alafuzoff et al., 2008), amyloid-beta plaques score 0–C, and Braak  $\alpha$ -synuclein stage 0–VI using the staging protocols of Brain Net Europe and Braak (Alafuzoff et al., 2009a, 2009b; Braak et al., 2006). The dissection of the caudate, putamen, hippocampus, middle frontal gyrus (F2), and middle temporal gyrus regions was performed from snap frozen human brain sections. Tissue was stored at  $-80^{\circ}\text{C}$  until further processing. Pathologic examination of the brain specimens showed changes consistent with the age of the individuals. The age at death, cause of death, and postmortem delay until dissection are provided in [Supplementary Table 1](#).

### 2.2. CAGE library preparation

Total RNA was extracted and purified from tissues using the Trizol tissue kit according to the instructions provided by the manufacturer (Invitrogen). RNA quality per library was assessed using the RNA integrity number with the Agilent Total RNA Nano kit (Table 1). The standard CAGE protocol (Kodzius et al., 2006) was adapted for sequencing on an Illumina platform. A thorough description of the protocol to prepare CAGE libraries and to sequence CAGE tags is presented in Takahashi et al. (2012). Briefly, complementary DNA (cDNA) was synthesized from total RNA using random primers, and this process was carried out at high temperature in the presence of trehalose and sorbitol to extend

**Table 1**  
Description of the tag counts per region/sample

Individual	Region	Batch <sup>a</sup>	RIN	Tag counts <sup>b</sup>	Unique counts <sup>c</sup>	Mapping rate <sup>d</sup>	Ribosome mapping <sup>e</sup>
<b>1</b>	<b>Caudate</b>	<b>1</b>	7.6	1,988,794	935,084	0.856	0.062
<b>1</b>	<b>Frontal</b>	<b>1</b>	7	3,453,682	1,531,751	0.866	0.049
<b>1</b>	<b>Hippocampus</b>	<b>1</b>	6.5	2,022,640	979,162	0.811	0.09
<b>1</b>	<b>Putamen</b>	<b>1</b>	7.7	3,814,753	1,627,659	0.826	0.069
<b>1</b>	<b>Temporal</b>	<b>1</b>	6.3	4,333,255	1,937,270	0.822	0.07
<b>2</b>	<b>Hippocampus</b>	<b>1</b>	6.5	1,682,943	310,481	0.843	0.072
<b>2</b>	<b>Caudate</b>	<b>2</b>	7.2	1,663,688	362,468	0.724	0.088
<b>2</b>	<b>Frontal</b>	<b>2</b>	6.9	1,745,155	801,757	0.822	0.04
<b>2</b>	<b>Putamen</b>	<b>2</b>	6.5	1,216,441	274,776	0.702	0.113
<b>2</b>	<b>Temporal</b>	<b>2</b>	6.8	936,396	259,968	0.748	0.103
<b>3</b>	<b>Frontal</b>	<b>2</b>	7.1	2,111,277	505,207	0.779	0.068
<b>3</b>	<b>Hippocampus</b>	<b>2</b>	8.8	1,785,386	413,336	0.816	0.041
<b>3</b>	<b>Temporal</b>	<b>2</b>	6.8	1,103,935	255,621	0.84	0.041
<b>4</b>	<b>Temporal</b>	<b>2</b>	5.9	1,199,974	356,840	0.71	0.127
<b>4</b>	<b>Frontal</b>	<b>2</b>	6.5	2,035,347	472,327	0.739	0.107
<b>4</b>	<b>Hippocampus</b>	<b>2</b>	6.4	1,251,589	335,644	0.731	0.109
<b>4</b>	<b>Putamen</b>	<b>2</b>	6.5	2,541,166	516,842	0.73	0.121
<b>5</b>	<b>Caudate</b>	<b>1</b>	7.9	3,096,524	1,144,105	0.875	0.059
<b>5</b>	<b>Putamen</b>	<b>1</b>	6.6	4,029,122	1,541,543	0.834	0.082
<b>6</b>	<b>Caudate</b>	<b>1</b>	7.4	3,587,220	1,296,765	0.875	0.053
<b>6</b>	<b>Putamen</b>	<b>1</b>	6.3	2,085,385	795,569	0.868	0.072
<b>7</b>	<b>Caudate</b>	<b>1</b>	6.8	4,875,578	1,625,317	0.862	0.062
<b>7</b>	<b>Frontal</b>	<b>2</b>	6.2	2,324,932	407,993	0.731	0.111
<b>7</b>	<b>Hippocampus</b>	<b>2</b>	6.2	3,158,604	597,669	0.775	0.033
<b>7</b>	<b>Temporal</b>	<b>2</b>	6.2	1,104,711	241,508	0.699	0.157

Details on the quality control and final counts used for the analysis are presented in [Supplementary data](#). Individual, region and batch id are presented in bold.

Key: RIN, RNA integrity number.

<sup>a</sup> Refers to 2 main batch effects corresponding to different period of times in which the cap analysis of gene expression libraries were prepared ([Supplementary data](#)).

<sup>b</sup> Refers to the total tag counts after removal of sequencing artifacts.

<sup>c</sup> Refers to the tag counts that map to single positions in the genome unique regions.

<sup>d</sup> Refers to proportion of tags that mapped to less than 10 positions.

<sup>e</sup> Refers to the proportion of tags that mapped to ribosomal DNA.

cDNA synthesis through GC-rich regions in 5' untranslated regions (UTRs). The 5' ends of messenger RNA within RNA-DNA hybrids were selected by the cap-trapper method (Kodzius et al., 2006) and ligated to a linker so that an EcoP15I recognition site was placed adjacent to the start of the cDNA, corresponding to the 5' end of the original messenger RNA. This linker was used to prime second-strand cDNA synthesis. Subsequent EcoP15I digestion released the 25- to 27-base pair (bp) CAGE tags. After ligation of a second linker, CAGE tags were polymerase chain reaction amplified, purified, and sequenced on the Illumina Genome Analyzer GLXII platform (Takahashi et al., 2012). The data have been submitted to the Gene Expression Omnibus (GEO) public repository (GSE43472).

### 2.3. DNA methylation microarrays

DNA isolation and purification to detect methylation was carried out following standard protocols (Supplementary data, Methods). Genome-wide amplified input and output samples were sent to Roche NimbleGen where they were hybridized to DNA Methylation 2.1 Million Deluxe Promoter Arrays. The arrays have a mean probe spacing of 99 bp and median probe spacing of 100 bp. Each array has more than 2.1 million probes distributed in the following manner (1) promoter regions from 7250 bases upstream of each TSS to 3250 bases downstream; (2) micro RNA (miRNA) genes, starting from 15 kbp upstream of the mature gene product to its 3' end; (3) CpG islands; and (4) ENCODE regions. Probes were chosen from the hg18 tiling database. Therefore, the probes targeted mainly annotated promoter regions and CpG islands.

### 2.4. Bioinformatics and statistical analysis of CAGE data

Primary quality control analysis included the removal of linker and barcode sequences as well as other sequencing artifacts to obtain raw CAGE tags of approximately 27 bps. Next, raw CAGE tags were mapped to the human genome (hg18 build) using Nexalign (T. Lassmann, <http://genome.gsc.riken.jp/osc/english/dataresource/>) allowing for 1 mismatch and 1 indel. The above steps were carried out with scripts and software (see Lassmann et al., 2009) developed at the RIKEN. Following previous approaches to analyze promoter activity based on CAGE data, we grouped raw CAGE tags into CAGE clusters using a clustering pipeline from Omics Science Center bioinformatics at the RIKEN (De Hoon et al., 2010). In brief, the CAGE tags that mapped to the same position in the human genome and were on the same strand were considered CAGE Transcription Start Sites (CTSSs) (level 1 [L1]). For tags that mapped to multiple positions in the genome, a rescuing approach was applied (Faulkner et al., 2008). L1 CAGE tags were clustered into level 2 tag clusters (L2 TCs) if they overlapped within 20 bps and were on the same strand. L2 TCs were grouped into level 3 (L3) TCs if they overlapped within a region of 400 bps and were on the same strand. For clarity, a CTSS marks the first nucleotide that is transcribed into RNA and is considered a putative TSS, whereas a L3 TC encompasses the region that is shared between proximal TSSs (Supplementary data, Methods) (Sandelin et al., 2007). After clustering approach, we obtained 6,735,699 CTSSs (L1 clusters). To increase the probability of capturing genuine promoter regions, we only selected L3 TCs that were present in at least 2 CAGE libraries and with a minimum count of 5 tags per million (TPM) (De Hoon et al., 2010) in at least 1 library; for example, only CTSS present at  $\geq 5$  TPM in one library and  $\geq 1$  TPM in another were included. For all downstream analysis, we used the L3 TCs. Unless stated otherwise, TCs refer to the L3 TCs.

Next, we annotated TCs to human genes by mapping the coordinates of the TCs to all available transcripts from GENCODE

version 3d. To do this, we downloaded all GENCODE transcripts from the UCSC genome browser (hg18 build; University of California, Santa Cruz, CA [UCSC]) at different levels of validation (<http://genome.ucsc.edu/cgi-bin/hgTables>). Custom Perl scripts and BEDTools (Quinlan and Hall, 2010) were used to map the coordinates of the TCs to genomic regions corresponding to specific transcriptional units (Carninci et al., 2006). TCs that did not map to a specific gene were considered intergenic. Further, we divided the TCs into mutually exclusive classes according to the gene region they mapped to. TCs that mapped to a 5' UTR or  $-300/+100$  bps of a known TSS (core promoter region) were labeled as canonical. The remaining noncanonical TCs were labeled as 5' UTR antisense, 3' UTR, 3' UTR antisense, intronic, exonic, intronic antisense, and exonic antisense.

We classified the genes to which the TCs mapped to according to the following Biotypes: protein-coding gene (if it had an open reading frame), long noncoding RNA (lncRNA), miRNA, pseudogene, processed transcript (no open reading frame, but transcribed and not classified into any other category), and other ncRNAs using the definitions from GENCODE (Harrow et al., 2006) ([http://www.gencodegenes.org/gencode\\_biotypes.html](http://www.gencodegenes.org/gencode_biotypes.html)).

#### 2.4.1. Differential gene expression and promoter usage derived from CAGE data across 5 brain regions

To obtain an overview of the expression (count) profile of the CAGE libraries, we first tested for differential expression across brain regions and subsequently identified patterns of differences between these regions by means of hierarchical clustering. We focused on autosomal TCs with a minimum of 9 tag counts per TC because this is the minimum number of counts needed to get reliable estimate of expression (Robinson et al., 2010). We built a model that takes into account both biological and technical variations, as we found that tag expression was subject to batch effects (Supplementary data, Results). The model assumes that CAGE tag counts ( $y_{ij}$ ) follow a negative binomial distribution, which is standard for modeling read/tag counts. It also includes brain group (5 levels corresponding to 5 regions), batch (2 levels corresponding to 2 main batches [Supplementary data, Results]), and individual (7 levels corresponding to 7 individuals) as covariates. Details of the mathematical and statistical procedure are presented in Supplementary data, Methods.

To identify differentially expressed TCs (DETC) showing similar differences among (a subset of) groups, we carried out hierarchical clustering (with Euclidean distance) based on the coefficients of brain regions, which are lower than 3 in absolute value. This was carried out with the R function `hclust` from package `stats` (with default agglomeration method). We chose the partition that maximized the average silhouette index width.

Functional enrichment analysis was subsequently done on clusters (modules) of DETCs using the PANTHER version 7.0 database (Supplementary data, Methods). All further functional pathway analyses were carried out using this database.

### 2.5. Bioinformatics and statistical analysis of methylation data

The log2 ratio of the probe intensity in the experimental sample against control DNA was determined. The log2 methylation signals were converted into methylation peaks (MPs) using the NimbleGen software (Roche) with default parameters (Supplementary data, Methods). Further, we removed MPs that mapped to X and Y chromosomes as well as those that overlapped with centromeres, telomeres, and segmental duplications. MPs overlapping with regions in which more than 1 segment was detected for a single sample were also removed. Next, we selected consensus MPs that were shared in a minimum of 2 samples. For this, we used the `plink`



software version 7 (Purcell et al., 2007) and identified shared methylated “segments” with the command: `plink file -segment-group`. Next, we used BEDTools (Quinlan and Hall, 2010) to map the MPs to annotated human genes (hg18) using GENCODE version 3d at different levels of annotation. We also mapped the MPs to CpG islands downloaded from UCSC browser (Fujita et al., 2011). Details of the experimental protocol and the downstream analysis are presented in [Supplementary data](#).

### 2.5.1. Differential methylation analysis

We modeled the log<sub>2</sub> ratios of the probe intensities taking both biological and technical variations into account and assuming that the ratios followed a normal distribution. Brain group (here we used the caudate as reference group), batch, and individual factors were covariates. We fitted 2 models per methylation probe: a full model, which included all 3 covariates and a null model where the factor brain group was discarded. We tested for differences in the models using a one-way analysis of variance, implemented in R version 13, and adjusted for multiple testing using the Bonferroni correction. Differentially methylated peaks (DMPs) were defined as differentially methylated probes occurring at a minimum overlap of 300 bps (R script provided by K. Lo at Roche, [k.lo@roche.com](mailto:k.lo@roche.com)).

### 2.5.2. Correlation between methylation signals and expression

First, we calculated the average methylation for every MP, adjusting for both biological and technical variations as mentioned previously. Next, we overlapped the genomic coordinates of the MPs with the genomic coordinates of the TCs (−1500/+500) using BEDTools (Quinlan and Hall, 2010) and estimated the Spearman correlation between the average mean intensity of methylation and the average expression of the overlapping TCs (geometric mean).

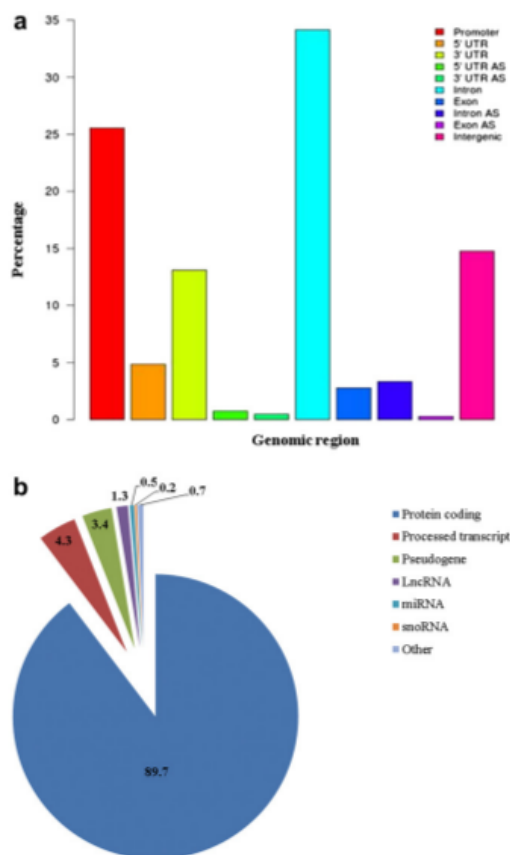
To test whether the expression of individual TCs were affected by methylation, we used the same statistical framework that we used to identify DETCs but included the methylation covariate as the variable of interest. Briefly, for each TC, we fitted 2 models. A full model with brain group, batch, and methylation as covariates, and a null model where methylation was removed. Because of the small number of MPs overlapping TCs, we could not fit the individual covariate. Significant differences were calculated as above.

## 3. Results

[Supplementary Fig. 1](#) shows a schema of the main steps of experimental procedure and the data analysis we carried out in this study. We prepared 25 CAGE libraries from total RNA isolated from the caudate nuclei, putamen, frontal and temporal lobes, and hippocampus from the 7 donors. In total, we sequenced 72 million CAGE tags (1–2 million per library approximately) in 5 sequencing rounds. [Table 1](#) summarizes the tag count and mapping rate per library after quality control ([Supplementary data, methods](#)). The final set of L3 TCs that were available for analysis numbered 70,202.

### 3.1. Features of brain transcriptome of aged individuals derived from CAGE

We mapped the TCs to 16,888 human genes from the GENCODE database (Raney et al., 2011). [Fig. 1a](#) shows that 31.2% of TCs mapped to the 5' UTR or promoter regions of previously annotated transcripts (canonical TCs), whereas the remaining 68.9% mapped to other regions including introns, exons, and 3' UTRs (noncanonical TCs). In addition, 13.6% of TCs did not map to any known transcript and were considered intergenic. Of these TCs, 559 (6%) mapped to lncRNAs (Jia et al., 2010) ([Supplementary Table 2](#)). Although canonical TCs represented less than one-third of all TCs ([Fig. 1a](#)), their expression was high and accounted for most of the overall TC



**Fig. 1.** Annotation of level 3 (L3) cap analysis of gene expression (CAGE) tag clusters (TCs) to human genes. (a) Barplot showing the percentage of TCs (y-axis) that map to different genomic regions: promoters, 5' untranslated regions (UTRs), 3' UTRs, antisense, introns, exons, antisense introns, antisense exons, antisense 5' UTR, antisense 3' UTR, and outside genes (intergenic). Promoter regions were defined as −300/100 base pairs relative to the 5' UTR. We defined canonical TCs those that mapped to promoters or 5' UTRs. The TCs that mapped to other regions were classified as noncanonical. The proportion of canonical TCs represents one-third of all TCs we identified. (b) Distribution of biotype classes for genes with canonical L3 CAGE TCs. Pie chart showing the percentage of genes with at least 1 canonical TC classified by biotype class: protein-coding genes (gene with open reading frame), long noncoding RNAs (lncRNAs), pseudogenes, micro RNAs (miRNA), small nucleolar RNAs (snoRNA), and processed transcripts (no open reading frame but transcribed and not classified into any other category).

expression. In contrast, the expression of most noncanonical TCs was low ([Supplementary Fig. 2](#)).

Of all the expressed genes, 14,479 (87%) had canonical TCs ([Supplementary data, Data set 1](#)). As shown in [Fig. 1b](#), 90% of these genes encode proteins. The remaining 10% consist of ncRNA, of which annotated pseudogenes account for 33%. We compared the list of genes that were expressed in our data set with those from RNASeq data from brain and other tissues (Ramskold et al., 2009). We found an overlap of 77% ([Supplementary Fig. 3](#)). Genes expressed in brain according to Ramskold et al. (2009) that were not present in our CAGE data set included both mitochondrial (e.g., *MT-ATP6*, *MT-ND3*, and *MT-CO2*) and ribosomal genes. In contrast,

there was a larger proportion of ncRNA in our brain CAGE data set (24% more compared with RNASeq, [Supplementary Fig. 3b](#)), with a particular enrichment for pseudogenes and lncRNAs.

We looked at the expression profile of 1909 highly expressed genes with canonical TCs (90th percentile of the log geometric mean of expression distribution; [Supplementary Table 3](#)) in more detail. This group included genes involved in brain aging (e.g., *GPAIP*, *SPARCL1*, and *B2M*, [Starkey et al., 2012](#)), calcium homeostasis (*CALM1–3*), neurodegeneration (*CLU* and *PICALM*, [Mengel-From et al., 2011](#)), and oxidative stress (e.g., *PTGD2*, *CA11* and *SOD1*, [Pareek et al., 2011](#)). We carried out functional enrichment analysis using PANTHER version 7.0 ([Mi et al., 2009](#); [Thomas et al., 2003](#)) on the group of highly expressed genes. Although many genes could not be classified, the most significant molecular pathways identified included the ubiquitin-proteasome pathway, synaptic transmission pathway, Huntington's disease, and PD ([Supplementary Fig. 4](#)). The overrepresentation of the PD pathway was mediated through genes encoding components of the ubiquitin-proteasome pathway (e.g., *PSMA1* and *PSMA2*), heat shock proteins (e.g., *HSPA2* and *HSPA5*), cell cycle components (e.g., *SEPT2*, *SEPT4*, and *SEPT5*), and synaptic genes (e.g., *SNCA*) among others. This shows that genes, for which mutations and/or variants that have been associated with PD, are components of cellular pathways that are highly expressed in the cortical and subcortical brain regions.

### 3.2. Extent of alternative promoter usage in brain transcriptome

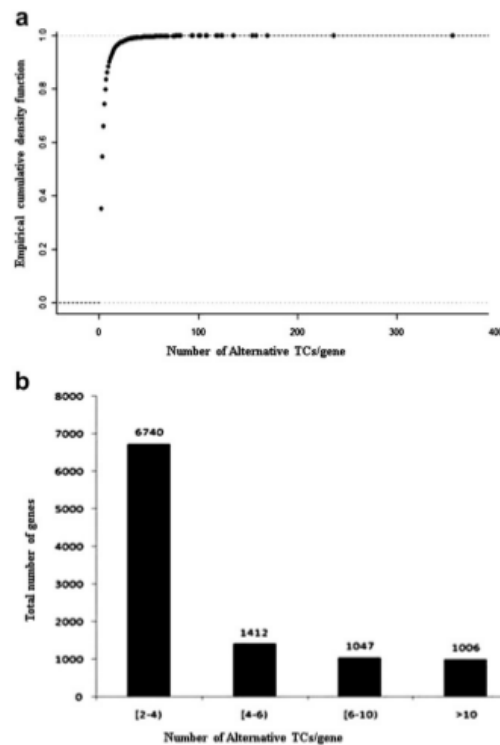
We defined alternative TCs (ATCs) as those that mapped to the same gene but were separated by a distance of >300 bp. TCs that were unique for a single gene were defined as “dominant TC” (DTC). Compared with DTCs, ATCs were mostly noncanonical and at least 34% of them mapped to introns.

In our data, 60% of genes (10,205 of 16,888 expressed genes) used ATCs (mean 5, range of 2–356, [Fig. 2](#)). Most genes with ATC had at least 1 canonical TC. We noted that the number of ATC per gene was above 10 for 10% of the genes ([Fig. 2](#)). Because some of the genes were quite large, we used linear regression to model the number of ATC per gene (for genes with at least 16 ATCs– 5% of the genes with large number of ATCs) against gene size. We found a correlation of about 0.3 ( $R = 0.28$ ,  $p$  value <2.2–16). This shows that gene length does not account to a large extent for the excess of ATC in genes. Outlier genes included *KCNIP4*, *PCDH9*, *CADM2*, *BAI3*, *NRG3*, *LSAMP*, *NRXN1*, *LRRTM4*, and *FGF14*, each with at least 100 ATCs. Functional enrichment analysis on genes with more than 16 ATCs (469 genes) showed an overrepresentation of glutamate receptor signaling and synaptic plasticity although most of the genes remained unclassified.

### 3.3. Regional differences in TC expression across the 5 brain regions

To identify signatures of gene expression across different brain regions, we sought CAGE clusters that were differentially expressed in one or more of the brain regions. We modeled the expression of the TCs using the number of counts and tested for significant differences in expression because of “regional effects” (see Section 2). We identified 7412 DETCs. Of these, 6037 were ATC of genes with a main canonical promoter. We identify neither any major differences in biotype between the differentially and nondifferentially expressed groups nor an excess of antisense TCs.

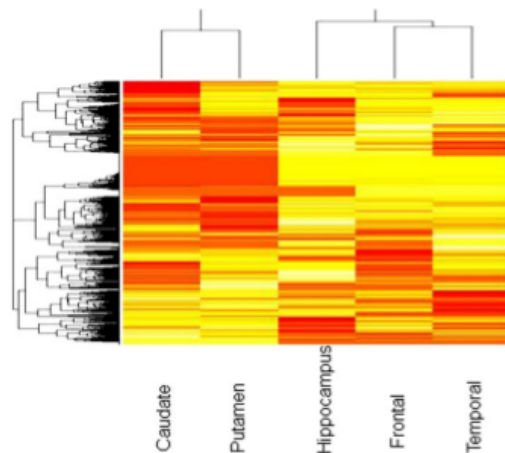
[Fig. 3](#) presents the results of the hierarchical clustering for the 7412 DETCs. We identified 3 main branches: one connecting the striatal regions (caudate and putamen), one connecting the cortical regions (frontal and temporal), and a third that separated the hippocampal region from the other 2 groups. [Fig. 3](#) also shows that the TCs were grouped into different clusters. We separated the



**Fig. 2.** Distribution of the number of alternative tag clusters (ATCs) per gene. (a) The empirical cumulative distribution (y-axis) of the number of ATCs per gene (x-axis) and (b) number of ATCs per bin category. The number of genes with 2 or more ATCs is shown at the top of every bin category.

DETCs into nonoverlapping modules (groups of TC that were differentially expressed in one or more regions) and identified 29 modules ([Table 2](#)). The largest module (M13) was characterized by small differences in expression across regions, and no region was clearly separated from the rest ([Supplementary Fig. 5a](#)). The other modules were characterized by more obvious differences in the average counts in 1 or 2 brain regions relative to the others ([Supplementary Fig. 5b–d](#)). These included M18 (lower expression in striatum vs. cortical regions and hippocampus), M4 (decreased expression in the caudate nucleus vs. the rest), M27 (lower expression in hippocampus), and M2 (increased expression in the cortex).

We evaluated whether specific signaling, metabolic, and disease pathways were enriched in the differentially expressed modules with at least 100 TCs. We used all the genes that were expressed in our data and that could be annotated in PANTHER version 7.0 as a reference set ([Supplementary Table 4a](#) shows the pathways that were significantly enriched in the reference group). Compared with the reference group, few pathways were enriched in the set of differentially expressed modules ([Supplementary Fig. 6](#)). The most significant pathway was the fibroblast growth factor (FGF) signaling pathway in M27 (lower expression in hippocampus) ( $p$  value <0.0005). Several genes from the FGF pathway were differentially expressed, including *FGF12*, *FGF14*, *RASA1*, *MAPK6*, *MAPK10*, *PPP2R2B*, and *PPA2*. All these genes had a main canonical TC that



**Fig. 3.** Unsupervised clustering of differentially expressed tag clusters (DETC). The graph depicts the unsupervised clustering of the  $\beta$  coefficients of the factor “region” derived from the statistical analysis of differences in expression because of regional effects (see Section 2). The dendrogram at the top shows that basal ganglia cluster together and that frontal and temporal cortices cluster together. The dendrogram at the left of the graph was used to split the DETCs into functional modules (see Results).

was uniformly expressed across brain regions and an ATC showing reduced expression in hippocampus. Other significant pathways ( $p$  value  $<0.005$ ) included platelet-derived growth factor signaling in M6 (lower expression in caudate compared with all other regions); synaptic trafficking in M2 (higher expression in cortex than in striatum and hippocampus) and M27 (lower expression in hippocampus); and glutamate receptor type I (metabotropic glutamate receptor group I [mGluRI]), *Wnt* signaling, and Huntington’s disease pathways in M18 (lower expression in striatum compared with cortex and hippocampus). These significant pathways mediate many cell functions including proliferation, differentiation, and survival (Goldbeter and Pourquie, 2008; Moon et al., 2004; Peng et al., 2010). A list of enriched pathways per module and genes with TC in each of the pathways is presented in [Supplementary Tables 4a and b](#), respectively.

#### 3.4. Unexpected expression of neurodevelopmental transcription factors in aged brain

To investigate whether differential promoter usage across brain regions can be explained by differences in the manner in which they are regulated, we searched for transcription factors (TFs) that were differentially expressed across the 5 regions. We mapped all DETCs to a manually curated list of TFs (Vaquerizas et al., 2009). We identified 519 DETCs that mapped to 320 TF genes, although only 20% mapped to the promoter or 5’ UTR region ([Supplementary Table 5](#)). The DETF with the highest expression included those involved in neuronal postmitotic differentiation and laminar integrity in the cortex (e.g., *TBR1*, [Bedogni et al., 2010](#), *NR2F1*, [Naka et al., 2008, \*NEUROD1\*, \*NEUROD2\*, \*BHLHE22\*, and \*MEF2C\*\) and neuronal plasticity \(e.g., \*NR4A1\*\) \(\[Table 3\]\(#\) presents the top 20 most highly expressed TF per module\). Most of the DETCs that mapped to TF were ATCs. One exception was a DTC that mapped to the promoter region of the \*KLF5\* gene and was differentially expressed in M27. \*KLF5\* has been shown to regulate survival and apoptosis](#)

through the regulation of MAPK kinase pathway. Other TFs that are module specific are presented in [Supplementary Table 5](#).

To identify specific TFs that were coexpressed with (and possibly regulate) the DETCs, we screened proximal ( $-300/+100$  bp) and distal ( $-1500/+500$  bp) promoter sequences of all TCs for transcriptional factor binding sites (TFBS) using remote dependency models (see [Supplementary data, Methods](#)). Overall, we identified 3 classes of TFBS that were significantly overrepresented in the promoter regions of the DETC, namely, BPTF (FAC1), the TBX family, and CUX1 (CDP). These TFs stand out as regulators during neurodevelopment including dendritogenesis (CUX1) ([Cubelos et al., 2010](#)), cortical formation (Tbr1-TBX) ([Bedogni et al., 2010](#)), and neurite outgrowth (BPTF) ([Rhodes et al., 2003](#)). On the other hand, we found that 15 classes of TFBS were significantly underrepresented including *E2F*, *EGR* (KROX), the Sp family (*Sp1* and *Sp3*), *Elk1*, *ATF6*, *CREB1*, and *MYC*, and *KLF5*. These TFs are known to be involved in apoptosis (*E2F* and *KLF5*) and synaptic plasticity (*EGR1-2*, *CREB*, *KLF5*, and *Elk*).

We also screened every module separately. We identified significant over/under-representation of TFBS in 19 out of the 29 modules ([Supplementary Table 6a and b](#)). The TBX binding site was overrepresented in most of the modules, whereas the BPTF binding site was significantly overrepresented in M13 and M27. Other TFBS were overrepresented although they did not reach statistical significance ([Supplementary Table 6a and b](#)).

#### 3.5. The extent of methylation in the brain transcriptome of aged individuals

DNA methylation at CpG nucleotides is another crucial mechanism for the regulation of gene expression ([Jones, 2012](#)). To investigate to what extent the patterns of expression in our data correlated with methylation, we analyzed methylation signatures in all 25 samples. After quality control and filtering, we obtained 551,178 MPs distributed and 95,715 of these were shared by at least 2 samples (of the 25 samples) and were used for downstream methylation analysis. We first assessed how many annotated genes from GENCODE were methylated and found that 73% of all methylation signals mapped within genes ([Fig. 4](#)), 43% to introns, 27% to exons, and 25% to promoter regions. We also looked at the proportion of methylation signals that occurred within CpG islands. We found that only 6% of methylated regions mapped within CpG islands. Of the promoters that mapped within CpG islands (45% of total), only 38% were methylated. Our data show that most of the methylated genomic regions occur in gene bodies and outside CpG islands (the list of MPs we used for the analysis is available on request).

##### 3.5.1. Regional differences in methylation across the 5 brain regions

To identify DMP, specific for specific brain regions, we modeled the MPs using a linear model for regional effects, adjusting for both individual and possible methylation batch effects. Using this approach, we identified 13,423 DMPs, and of these 75.9% were mapped within gene bodies. Genes that were differentially methylated included *NRXN1*, *ITPR1*, *MADD*, *CNTNAP1*, *SRR*, *GABBR1*, *INPP5A*, *HTR1D*, *DLGAP1*, and *TIAM2*, which have been previously shown as methylated ([Iwamoto et al., 2011](#)) and that we found differentially methylated in frontal cortex.

We also compared the list of DMPs with MPs derived from [Davies et al. \(2012\)](#), where differences in methylation across several brain areas (mainly cortex and cerebellum) and blood were reported. We found that at least 39% of the DMPs overlapped with these from [Davies et al. \(2012\)](#). Moreover, several genes that we found differentially methylated showed also differences in methylation between cerebellum and cortex (e.g., *AACS*, *ADCY5*, *EPHB4*,



**Table 2**  
Number of DE clusters identified for the DETCs

Module id	No. TCs	Caudate	Putamen	Hippocampus	Frontal	Temporal	Proportion of all DE TC
13	3190						0.43
18	1063						0.143
6	683						0.092
27	295						0.04
2	273						0.037
20	256						0.035
23	170						0.023
4	163						0.022
19	164						0.022
10	155						0.021
16	119						0.016
8	116						0.016
11	107						0.014
3	107						0.014
9	91						0.012
7	74						0.01
22	51						0.007
1	41						0.006
25	45						0.006
17	43						0.006
5	38						0.005
12	39						0.005
26	38						0.005
15	28						0.004
28	12						0.002
21	18						0.002
29	12						0.002
24	11						0.001
14	10						0.001
<b>Total</b>	<b>7412</b>						

Dark gray represents higher expression relative to other regions. Light gray represents lower expression relative to other regions. Black represents similar expression profile for all regions.

Key: DETCs, differentially expressed tag clusters.

*GALNT9*, and *GRM4*) and between brain and blood (e.g., *CCDC85A*, *PCDH9*, *PDE4D*, and *PPP2R2B*). This analysis shows that as much as 39% of methylated regions in brain (as identified by 2 different approaches) exhibit differences in their methylation profile in the brain regions we analyzed. The list of DMPs that we identified and that overlapped with MPs from Davies et al. (2012) is presented in Supplementary data, Data set 2).

### 3.5.2. Correlation between MPs and expression

To analyze the correlation between expression and methylation in our data, we first overlapped the genomic coordinates of both data sets considering promoter regions from −1500 to +500 bp relative to the most highly expressed TSS. We found that only 9%

of all TCs overlapped with at least 1 MP. Overall, there was no significant correlation between methylation and expression (Spearman correlation:  $r = -0.05$ ), most likely because of the large variation in the methylation of TCs with very low counts (Supplementary Fig. 8). We also analyzed the correlation between methylation and expression for protein-coding genes and non-coding genes separately (the number of ncRNA genes that overlapped with the MPs was too small to be analyzed independently) and did not observe any difference in their correlation coefficients (Spearman correlation of −0.06 and −0.07 for ncRNAs and protein-coding genes, respectively). Therefore, we tested for significant differences in expression because of “methylation effects” at individual TCs adjusting the expression for brain region and batch

**Table 3**  
List of 20 most highly differentially expressed TF

TC ID	Start	End	TF	Module	Mean (geometric)
L3_chr2_+161981068	161980893	161981527	TBR1 ( <i>tbx family</i> )	13	28.718
L3_chr5_+92946017	92945793	92946068	NR2F1( <i>COUP-1f1</i> )	13	6.658
L3_chr12_+50731491	50731420	50731653	NR4A1	13	6.035
L3_chr7_+39092007	39091721	39092121	POU6F2	13	5.520
L3_chr8_+65655474	65655301	65655790	BHLHE22	13	5.035
L3_chr19_+41561943	41561901	41561975	ZFP14	13	4.983
L3_chr5_+88155431	88155327	88155565	MEF2C	13	4.965
L3_chr1_+925340	925274	925452	HES4	13	4.738
L3_chr2_+182253487	182253446	182253729	NEUROD1	13	4.271
L3_chr2_+242205564	242205419	242205632	THAP4	13	3.994
L3_chr17_+35017699	35017598	35017742	NEUROD2	13	3.946
L3_chr3_+69871321	69871264	69871369	MITF	13	3.937
L3_chr13_+72531139	72531098	72531259	KLF5	27	3.847
L3_chr4_+146623601	146623337	146623645	SMAD1	13	3.750
L3_chr9_+37455447	37455266	37455461	ZBTB5	13	3.745
L3_chr13_+73606569	73606482	73606578	KLF12	13	3.587
L3_chr1_+13977672	13977542	13977772	PRDM2	13	3.446
L3_chr19_+60846825	60846530	60846828	ZNF581	13	3.399
L3_chr7_+38984037	38983927	38984054	POU6F2	13	3.380
L3_chr2_+45022343	45022302	45022747	SDX3	3	3.366

Key: TC, tag cluster; TF, transcription factor.

covariates. For this analysis, we only considered MPs that were present in at least 5 libraries. After correcting for multiple testing, we identified 312 TCs (5%) with differences in expression because of methylation effects. Of these, 34 TCs also exhibited differences in expression per region. Therefore, the differential expression because of regional effects we observed earlier was not driven by differences in methylation to a large extent. Genes with differences in expression per region because of methylation status included *CDK10*, *NRN1*, *PYCARD*, *TIMP3*, and *UCP2*. For these genes, promoter methylation has previously shown to regulate expression (Gloss et al., 2011; Konishi et al., 2011).

3.6. Effects of methylation on TF expression and TFBS

We compared the methylation status of differentially and non-DETFs. We did not identify any significant difference in the proportion of methylated TFs between the 2 groups (7% and 10% for differentially and non-DETFs, respectively). However, there was

a significantly higher proportion of MPs mapping to the 3' UTR regions in the DETFs (50% vs. 15%, Fisher  $p = 0.0001$ ), whereas in the group of non-DETFs, most of the MPs mapped to the canonical promoter region (11% vs. 42%, Fisher  $p = 0.0058$ ).

We also analyzed whether methylation could affect the expression of TCs by binding to their TFBS, presumably by modifying the spatial structure of binding sites (Choy et al., 2010). We screened the TFBS identified previously for overlaps with differentially MPs and found 304 TFBS in such locations (details of the statistical analysis are provided in Supplementary data, Methods). Out of all these TFBS, we only selected those, which overlap with differentially MPs showing a negative correlation between expression and methylation. Because of low number of high confident TFBS predictions made by RDM, we only identified a few TFs having several binding sites in such locations, namely, E2F group, Sp1:Sp3 complex, AP2alphaA, FAC1, and NHLH1 (for details, see Supplementary data, Methods and Table 7). This coincided with the underrepresentation of predicted TFBS for certain TFBS

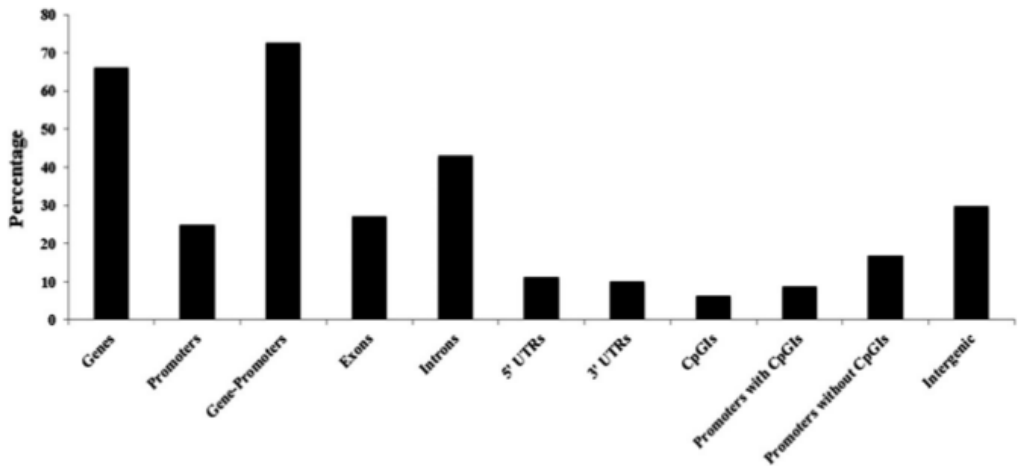


Fig. 4. Percentage of methylation peaks mapping to different gene regions, within and outside CpG islands.

including E2F and Sp1:Sp3 observed earlier (Supplementary Table 6a and b), suggesting that the corresponding TFs even being nondifferentially expressed may be involved in regulation of differential expression.

#### 4. Discussion

In this study, we used CAGE in combination with massive parallel sequencing to profile transcription initiation across 5 different brain regions of aged, nondemented individuals and evaluated the extent of region specificity in alternative promoter usage and expression. At a sequencing depth of 1–2 million CAGE tags per library, we found that 40% of all GENCODE genes were expressed in brain. This estimate is probably conservative because it has been shown that deeper sequencing is needed to identify rare functional transcripts (Mercer et al., 2011). In addition, we annotated 6% of intergenic TCs to 559 lncRNAs that had previously only been predicted *in silico*.

We found that 77% of the genes with canonical TCs in our data set overlap with another brain transcriptome data set derived from RNA-Seq methodology (Ramskold et al., 2009). Comparing the 2 data sets reveals that CAGE detects more ncRNA transcripts (e.g., lncRNAs and pseudogenes) whereas the proportion of protein-coding genes was higher with RNA-Seq. These differences could be the result of differences in sequencing depth or due to marked differences in the experimental design of both approaches. Indeed, although CAGE and RNA-Seq can be used to quantify the amount of gene expression and that there is a high correlation of gene expression between these 2 approaches (0.57, see Dong et al., 2012), RNA-Seq libraries are commonly enriched for poly A+ transcripts (Mortazavi et al., 2008) of which protein-coding genes are an abundant class. In contrast, CAGE method captures capped RNA transcripts of both poly A+ and poly A– classes (Carninci, 2007). This also may explain why some genes that appeared highly expressed in brain in the RNA-Seq data set were not identified with CAGE including mitochondrial and ribosomal genes because they are uncapped and, therefore, not well covered by the CAGE approach.

Recent studies show that ncRNAs regulate gene expression in brain and play a role in the development and in the onset of neurologic diseases (Schonrock et al., 2011). Most research has focused on deciphering the functional role of lncRNAs and miRNAs, but other classes of ncRNAs may also be important. We found that more than 4.7% of the total RNA pool (and 24% of the ncRNA) consisted of annotated pseudogenes. The contribution of this ncRNA class to the transcriptome is currently unknown, with estimates ranging from 5% (Frith et al., 2006), which is consistent with our data, to 20% (Pink et al., 2011). Although the functional impact of ncRNA classes was not assessed in this study, our findings demonstrate that pseudogene expression is a pervasive feature of the transcriptome in aged brain.

We found expression patterns consistent with aging, including high expression of *GPAFP* (Starkey et al., 2012) and *SPARCL1*, which are markers of gliosis, and high expression of genes involved in protection against oxidative stress and amyloid aggregation. This group includes *CLU*, the gene for clusterin, an extracellular chaperone that maintains stressed proteins in a soluble state, thereby preventing their precipitation (Poon et al., 2002). Clusterin colocalizes with amyloid plaques and neurofibrillary tangles, and it has been suggested that it protects neurons from aggregate-induced damage (Yerbury et al., 2007). The ubiquitin-proteasome pathway was overrepresented in the group of highly expressed genes. This pathway has been shown to be downregulated in disorders such as AD and PD (Dennissen et al., 2012), and this decrease correlates with a failure of neurons to remove toxic protein aggregates. In this

regard, it is important to stress that despite some pathologic findings consistent with aging, none of the 7 donors used for this study showed any overt AD or PD pathology (Braak tangle stages  $\leq 3$  and Braak  $\alpha$ -synuclein stage 0–IV; Supplementary Table 1). These results suggest that increased expression of genes involved in the ubiquitin-proteasome pathway and neuroprotection (e.g., *CLU*) may help to protect against overt protein aggregation in aged healthy individuals.

It has been recently shown that alternative promoter usage and alternative splicing can explain differences in gene expression across brain regions (Pal et al., 2011; Tollervey et al., 2011). Our data support the role of alternative promoters in causing expression differences between brain regions. We found that 81% of the DETCs were putative alternative TSS of genes with a main promoter that was similarly expressed in all the regions analyzed. This shows that the major transcripts were more often uniformly expressed whereas alternative transcripts were more likely to be region specific. Alternative promoters can alter the expression of a main transcript by competing for the cell's transcription machinery (Davuluri et al., 2008) or by antagonizing the effects of the main transcript (Tschan et al., 2003). For example, we found a DETC in M18 (Supplementary Table 5) mapping to the promoter region of a short isoform of *DMTF1*, which has been shown to antagonize the effects of the main *DMTF1* transcript in myeloid lines (Tschan et al., 2003). Whether the expression of the shorter isoform leads to the same changes observed in other cells cannot be ascertained here, but it suggests an interesting mechanism by which alternative promoter usage might lead to differences in expression.

In our data, most of the ATCs that were differentially expressed were located in noncanonical gene regions (Fig. 1a), particularly in introns. Although there is evidence that CAGE tags can also mark post-transcriptional events (Mercer et al., 2010), we provide several lines of evidence indicating that a proportion of transcription is initiated from noncanonical gene regions. First, we only included CAGE clusters present in at least 2 biological replicas, which makes it unlikely that a tag identified twice is the result of an artifact. Second, we found that at least one-third of noncanonical TCs overlapped with other signatures of promoter activity derived from H3K4me3 histone marks (data not shown). In addition, we confirmed with RACE the existence of capped products for 4 putative alternative TSSs in the *CNP*, *RTN4*, *NRG3*, and *AUTS2* genes (Supplementary data, Results), which may represent novel isoforms for those genes. Indeed, we confirmed experimentally the presence of an alternative TSS in the intronic region of *AUTS2*, which is associated with a shorter transcript that was previously only *in silico* predicted. Our results indicate that at least one-third of alternative TSS map to intronic gene regions.

Several growth factor signaling pathways have been implicated in the alterations that render neuronal cell populations susceptible to neurodegeneration. Our data showed that the *FGF*, epidermal growth factor (*EGF*), insulin growth factor (*IGF*), and platelet-derived growth factor pathways were overrepresented in several differentially expressed modules (Supplementary Fig. 5 and Table 4a). Common to these pathways is the mitogen-activated protein kinase (MAPK) cascade that has a broad range of effects on cellular function including survival and differentiation (Thomas and Huganir, 2004). The FGF signaling pathway was the most significantly overrepresented pathway in module M27, where a reduced expression in hippocampus was observed. The hippocampal region is a primary target of the neurodegenerative changes that lead to cognitive impairment and AD. Several mechanisms have been suggested to lead to hippocampal dysfunction, including decreased neuronal plasticity and increased calcium toxicity. The FGF pathway can influence neural plasticity through several mechanisms including MAPK/ERK activation (Thomas and Huganir, 2004), and



its expression was reduced in the hippocampus relative to other regions. These findings suggest that the FGF pathway could be an important target for pharmacologic treatments to combat neurodegeneration.

The caudate and putamen regions (striatum), which are components of the cortical-subcortical circuits of motor functions, are particularly susceptible to neurodegeneration in disorders such as Huntington's disease and PD (DeLong and Wichmann, 2007). Interestingly, functional enrichment analysis based on several DETCs showed that genes encoding components of the *Wnt* signaling pathway and the mGluRI were significantly over-represented in the modules where coexpression in the striatal regions was observed (M18; Table 2). Both *Wnt* signaling and mGluRI have been implicated in the development or progression of PD (Johnson et al., 2009; L'Episcopo et al., 2011). Moreover, mGluRI modulates neurotransmission throughout the basal ganglia, and its deregulation can contribute to neuronal damage (Johnson et al., 2009). Our results suggest that in the absence of a clear genetic risk, pathways other than those associated with classical mutations are important determinants of the regional vulnerability in the aging brain.

We investigated whether differences in expression could be attributed to differential TF expression. We found that 7% of TFs were differentially expressed, and many of these have been shown to be involved in the neurodevelopment, which is unexpected given that neurons are postmitotic cells. The TFBS analysis also showed an overrepresentation binding sites for TFs involved in neurodevelopment. There are few explanations for this finding including a bias in the literature toward functional annotation of neurodevelopmental TFs. Another plausible explanation is that, as the brain ages, these genes may become derepressed because of, for example, damage in their promoter regions. Although we did not find decreased methylation in the group of DETFs, we found decreased methylation in the promoter region of this group and increased methylation in the 3' UTRs. Methylation marks at both ends of transcriptional units could affect the expression of the group of DETFs (Jones, 2012).

Our analysis of methylation indicated that most of the methylation signals in our samples mapped to gene bodies and outside CpG islands. This is consistent with recent evidence that in brain most methylation signals occur within gene bodies, most likely in association with alternative promoters (Maunakea et al., 2010). However, we did not find an overall correlation between methylation and TC expression. Several factors could account for the lack of correlation. For example, batch effects were evident in the CAGE data set. In addition, only 9% of the methylated regions colocalized with a TC, which means that most of the expression in our data remained uninvestigated. The lack of overlap between the MPs and the CAGE clusters could also be because of the fact that the arrays we used to profile methylation were biased toward annotated promoters and CpG islands, whereas our CAGE clusters mapped to a large extent to noncanonical regions. Last, as a result of the small sample size, most of MPs were identified in less than 5 samples and were removed from the statistical analyses. Despite this drawback, we identified several gene-associated TCs that were affected by methylation, some of which were already documented (Iwamoto et al., 2011).

Our study is far from being comprehensive because of our small sample size and the limited number of brain regions analyzed. In addition, because of the diverse cellular composition of the brain, one might argue that the expression we observed is not exclusive to neuronal populations, although neurons and glia cells represent most of the cellular pool in human brain. A separate issue is that most of our bioinformatics analysis used public databases, which are still incomplete. For example,

many protein-coding genes that we found differentially expressed could not be assigned to any functional pathway because of a lack of annotation. Therefore, inferences about functional pathways are based on a limited number of genes. Nonetheless, our data set provides an important addition to existing data on spatial expression patterns in brain.

In summary, our study shows that despite the absence of neuropathologic hallmarks of neurodegenerative disease, genetic signatures related to neurodegeneration were already present in brain regions that are highly vulnerable to neurologic disorders. We showed that differences in transcription initiation and hence gene expression between brain regions are partly explained by alternative promoter usage and that specific signaling pathways are affected by the differential patterns in gene expression that we observed. Our data are a starting point to investigate regional susceptibility to brain aging and neurodegeneration.

#### Disclosure statement

The authors declare no conflicts of interest.

#### Acknowledgements

The authors thank the Netherlands Brain Bank (Amsterdam, the Netherlands), especially Michiel Kooreman, for providing excellent postmortem human brain tissue and are grateful to all controls who donated their brains, K. Lo (Roche) for providing with an R script to identify differentially methylated regions, and Aad W. van der Vaart and Wessel N. van Wieringen for useful discussions on the statistical analysis of differential expression. G.G.R.L. is partly supported by the Center for Medical Systems Biology, established by the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research. M.F. is funded by Portuguese Foundation for Science and Technology (scholarship reference SFRH/BD/33536/2008). J.S. is funded by Hersenstichting Nederland Fellowship project B08.03 and the Neuroscience Campus Amsterdam.

#### Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.neurobiolaging.2013.01.005>.

#### References

- Abdolmaleky, H.M., K-h. Cheng, Russo, A., Smith, C.L., Faraone, S.V., Wilcox, M., Shafa, R., Glatt, S.J., Nguyen, G., Ponte, J.F., Thiagalingam, S., Tsuang, M.T., 2005. Hypermethylation of the reelin (RELN) promoter in the brain of schizophrenic patients: a preliminary report. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 134B, 60–66.
- Alafuzoff, I., Arzberger, T., Al-Sarraj, S., Bodi, I., Bogdanovic, N., Braak, H., Bugiani, O., Del-Tredici, K., Ferrer, I., Gelpi, E., Giaccone, G., Graeber, M.B., Ince, P., Kamphorst, W., King, A., Korkolopoulou, P., Kovacs, G.G., Larionov, S., Meynert, D., Monoranu, C., Parchi, P., Patsouris, E., Roggendorf, W., Seilhean, D., Tagliavini, F., Stadelmann, C., Streichenberger, N., Thal, D.R., Wharton, S.B., Kretschmar, H., 2008. Staging of neurofibrillary pathology in Alzheimer's Disease: a study of the BrainNet Europe Consortium. *Brain Pathology* 18, 484–496.
- Alafuzoff, I., Ince, P.G., Arzberger, T., Al-Sarraj, S., Bell, J., Bodi, I., Bogdanovic, N., Bugiani, O., Ferrer, I., Gelpi, E., Gentleman, S., Giaccone, G., Ironside, J.W., Kavantzis, N., King, A., Korkolopoulou, P., Kovacs, G.G., Meynert, D., Monoranu, C., Parchi, P., Parkkinen, L., Patsouris, E., Roggendorf, W., Roemmler, A., Stadelmann-Nessler, C., Streichenberger, N., Thal, D.R., Kretschmar, H., 2009a. Staging/typing of Lewy body related alpha-synuclein pathology: a study of the BrainNet Europe Consortium. *Acta neuropathologica* 117, 635–652.
- Alafuzoff, I., Thal, D.R., Arzberger, T., Bogdanovic, N., Al-Sarraj, S., Bodi, I., Boluda, S., Bugiani, O., Duyckaerts, C., Gelpi, E., Gentleman, S., Giaccone, G., Graeber, M., Hortobagyi, T., Hofberger, R., Ince, P., Ironside, J.W., Kavantzis, N., King, A., Korkolopoulou, P., Kovacs, G.G., Meynert, D., Monoranu, C., Nilsson, T., Parchi, P., Patsouris, E., Pikkarainen, M., Revesz, T., Roemmler, A., Seilhean, D.,

- Schulz-Schaeffer, W., Streichenberger, N., Wharton, S.B., Kretschmar, H., 2009b. Assessment of beta-amyloid deposits in human brain: a study of the BrainNet Europe Consortium. *Acta neuropathologica* 117, 309–320.
- Bandopadhyay, R., Kingsbury, A.E., Cookson, M.R., Reid, A.R., Evans, I.M., Hope, A.D., Pittman, A.M., Lashley, T., Canet-Aviles, R., Miller, D.W., McLendon, C., Strand, C., Leonard, A.J., Abou-Sleiman, P.M., Healy, D.G., Ariga, H., Wood, N.W., de Silva, R., Revész, T., Hardy, J.A., Lees, A.J., 2004. The expression of DJ-1(PARK7) in normal human CNS and idiopathic Parkinson's disease. *Brain* 127, 420–430.
- Bedogni, F., Hodge, R.D., Elsen, G.E., Nelson, B.R., Daza, R.A.M., Beyer, R.P., Bammler, T.K., Rubenstein, J.L.R., Hevner, R.F., 2010. Tbr1 regulates regional and laminar identity of postmitotic neurons in developing neocortex. *Proc. Natl. Acad. Sci. U S A* 107, 13129–13134.
- Braak, H., Bohl, J.R., Müller, C.M., Rub, U., de Vos, R.A., Del Tredici, K., 2006. Stanley Fahn Lecture 2005: the staging procedure for the inclusion body pathology associated with sporadic Parkinson's disease reconsidered. *Mov. Disord.* 21, 2042–2051.
- Caminci, P., 2007. Constructing the landscape of the mammalian transcriptome. *J. Exp. Biol.* 210, 1497–1506.
- Caminci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V.B., Brenner, S.E., Batalov, S., Forrest, A.R., Zavolan, M., Davis, M.J., Wilming, L.G., Aidinis, V., Allen, J.E., Ambesi-Impombato, A., Apweiler, R., Aturaliya, R.N., Bailey, T.L., Bansal, M., Baxter, L., Beisel, K.W., Bersano, T., Bono, H., Chalk, A.M., Chiu, K.P., Choudhary, V., Christoffels, A., Clutterbuck, D.R., Crowe, M.L., Dalla, E., Dalrymple, B.P., De Bono, B., Della Gatta, G., di Bernardo, D., Down, T., Engstrom, P., Fagioli, M., Faulkner, G., Fletcher, C.F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T.R., Gojbori, T., Green, R.E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T.K., Hirokawa, N., Hill, D., Huminecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S.P., Kruger, A., Kummerfeld, S.K., Kurochkin, I.V., Lareau, L.F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakaguchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K.C., Pavan, W.J., Pavese, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J.F., Ring, B.Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S.L., Sandelin, A., Schneider, C., Schonbach, C., Sekiguchi, K., Sempke, C.A., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugita, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S.L., Tang, S., Taylor, M.S., Tegner, J., Teichmann, S.A., Ueda, H.R., van Nimwegen, E., Verardo, R., Wei, C.L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S.M., Teasdale, R.D., Liu, E.T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J.S., Hume, D.A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, K., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y., 2005. The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563.
- Caminci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Sempke, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C., Forrest, A.R., Alkema, W.B., Tan, S.L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesl, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S.M., Wells, C.A., Orlando, V., Wahlestedt, C., Liu, E.T., Harbers, M., Kawai, J., Bajic, V.B., Hume, D.A., Hayashizaki, Y., 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature genetics* 38, 626–635.
- Choy, M.-K., Movassagh, M., Goh, H.-G., Bennett, M., Down, T., Foo, R., 2010. Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated. *BMC Genomics* 11, 519–529.
- Cubelos, B., Sebastián-Serrano, A., Beccari, L., Calkagnotto, M.E., Cisneros, E., Kim, S., Dopazo, A., Alvarez-Dolado, M., Redondo, J.M., Bovolenta, P., Walsh, C.A., Nieto, M., 2010. Cux1 and Cux2 regulate dendritic branching, spine morphology, and synapses of the upper layer neurons of the cortex. *Neuron* 66, 523–535.
- Davies, M.N., Volta, M., Pidsley, R., Lunnon, K., Dixit, A., Lovestone, S., Coarfa, C., Harris, R.A., Milosavljevic, A., Troakes, C., Al-Sarraj, S., Dobson, R., Schalkwyk, L.C., Mill, J., 2012. Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biol.* 13, R43.
- Davuluri, R.V., Suzuki, Y., Sugano, S., Plass, C., Huang, T.H., 2008. The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet.* 24, 167–177.
- De Hoon, M.J., Bertin, N., Chalk, A.M., 2010. Using CAGE data for quantitative expression. In: Caminci, P. (Ed.), *Cap Analysis Gene Expression (CAGE): The Science of Decoding Gene Transcription*. Pan Stanford Publishing Pte. Ltd, Yokohama, pp. 101–121.
- DeLong, M.R., Wichmann, T., 2007. Circuits and circuit disorders of the basal ganglia. *Arch. Neurol.* 64, 20–24.
- Dennis, F.J., Kholod, N., van Leeuwen, F.W., 2012. The ubiquitin proteasome system in neurodegenerative diseases: culprit, accomplice or victim? *Progress in neurobiology* 96, 190–207.
- Dong, X., Greven, M.C., Kundaje, A., Djebali, S., Brown, J.B., Cheng, C., Gingeras, T.R., Gerstein, M., Guigó, R., Birney, E., Weng, Z., 2012. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* 13, R53.
- Double, K.L., Halliday, G.M., Kril, J.J., Harasty, J.A., Cullen, K., Brooks, W.S., Creasey, H., Broe, G.A., 1996. Topography of brain atrophy during normal aging and Alzheimer's disease. *Neurobiology of aging* 17, 513–521.
- Double, K.L., Reyes, S., Werry, E.L., Halliday, G.M., 2010. Selective cell death in neurodegeneration: why are some neurons spared in vulnerable regions? *Progress in neurobiology* 92, 316–329.
- Faulkner, G.J., Forrest, A.R., Chalk, A.M., Schroder, K., Hayashizaki, Y., Carninci, P., Hume, D.A., Grimmond, S.M., 2008. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics* 91, 281–288.
- Frith, M.C., Wilming, L.G., Forrest, A., Kawai, H., Tan, S.L., Wahlestedt, C., Bajic, V.B., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Bailey, T.L., Huminecki, L., 2006. Pseudo-messenger RNA: phantoms of the Transcriptome. *PLoS genetics* 2, e23.
- Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T.R., Gardine, B.M., Harte, R.A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R.M., Leamed, K., Li, C.H., Meyer, L.R., Pohl, A., Raney, B.J., Rosenbloom, K.R., Smith, K.E., Haussler, D., Kent, W.J., 2011. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* 39, D876–D882.
- Gloss, B.S., Patterson, K.J., Barton, C.A., Gonzalez, M., Scurry, J.P., Hacker, N.F., Sutherland, R.L., O'Brien, P.M., Clark, S.J., 2011. Integrative genome-wide expression and promoter DNA methylation profiling identifies a potential novel panel of ovarian cancer epigenetic biomarkers. *Cancer Letters* 318, 76–85.
- Goldbeter, A., Pourquie, O., 2008. Modeling the segmentation clock as a network of coupled oscillations in the Notch, Wnt and FGF signaling pathways. *Journal of Theoretical Biology* 252, 574–585.
- Hardy, J., Selkoe, D.J., 2002. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science* 297, 353–356.
- Harrow, J., Denoeud, F., Frankish, A., Reynolds, A., Chen, C.-K., Chrast, J., Lagarde, J., Gilbert, J., Storey, R., Swarbreck, D., Rossier, C., Ucla, C., Hubbard, T., Antonarakis, S., Guigo, R., 2006. GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* 7 (Suppl 1), S4.1–S4.9.
- Iwamoto, K., Bundo, M., Ueda, J., Oldham, M.C., Ukai, W., Hashimoto, E., Saito, T., Geschwind, D.H., Kato, T., 2011. Neurons show distinctive DNA methylation profile and higher interindividual variations compared with non-neurons. *Genome Res.* 21, 688–696.
- Jia, H., Osak, M., Bogu, G.K., Stanton, L.W., Johnson, R., Lipovich, L., 2010. Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA* 16, 1478–1487.
- Joachim, C.L., Mori, H., Selkoe, D.J., 1989. Amyloid beta-protein deposition in tissues other than brain in Alzheimer's disease. *Nature* 341, 226–230.
- Johnson, K.A., Conn, P.J., Niswender, C.M., 2009. Glutamate receptors as therapeutic targets for Parkinson's disease. *CNS & neurological disorders drug targets* 8, 475–491.
- Johnson, M.B., Kawasawa, Y.I., Mason, C.E., Krnsnik, Z., Coppola, G., Bogdanovic, D., Geschwind, D.H., Mane, S.M., State, M.W., Sestan, N., 2009. Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* 62, 494–509.
- Jones, P.A., 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews* 13, 484–492.
- Kang, H.J., Kawasawa, Y.I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A.M.M., Pletikos, M., Meyer, K.A., Sedmak, G., Guennel, T., Shin, Y., Johnson, M.B., Krnsnik, Z., Mayer, S., Furtuzinhos, S., Umlauf, S., Ligo, S.N., Vortmeyer, A., Weinberger, D.R., Mane, S., Hyde, T.M., Huttner, A., Reimens, M., Kleinman, J.E., Sestan, N., 2011. Spatio-temporal transcriptome of the human brain. *Nature* 478, 483–489.
- Khaitovich, P., Muetzel, B., She, X., Lachmann, M., Hellmann, I., Dietzsch, J., Steigle, S., Do, H.H., Weiss, G., Enard, W., Heissig, F., Arendt, T., Niesek-Struwe, K., Eichler, E.E., Paabo, S., 2004. Regional patterns of gene expression in human and chimpanzee brains. *Genome research* 14, 1462–1473.
- Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y., Carninci, P., 2006. CAGE: cap analysis of gene expression. *Nat. Meth.* 3, 211–222.
- Konishi, K., Watanabe, Y., Shen, L., Guo, Y., Castoro, R.J., Kondo, K., Chung, W., Ahmed, S., Jelinek, J., Bumber, Y.A., Estecio, M.R., Maegawa, S., Kondo, Y., Itoh, F., Imawari, M., Hamilton, S.R., Issa, J.-P., 2011. DNA methylation profiles of primary colorectal carcinoma and matched liver metastasis. *PLoS one* 6, e27889.
- Konopka, G., Geschwind, D.H., 2010. Human brain evolution: harnessing the genomics (r)evolution to link genes, cognition, and behavior. *Neuron* 68, 231–244.
- L'episcopo, F., Serapide, M.F., Tirollo, C., Testa, N., Caniglia, S., Morale, M.C., Pluchino, S., Marchetti, B., 2011. Wnt1 regulated Frizzled-1/beta-Catenin signaling pathway as a candidate regulatory circuit controlling mesencephalic dopaminergic neuron-astrocyte crosstalk: therapeutic relevance for neuron survival and neuroprotection. *Molecular Degeneration* 6, 49–78.
- Lassmann, T., Hayashizaki, Y., Daub, C.O., 2009. TagDust—a program to eliminate artifacts from next generation sequencing data. *Bioinformatics* 25, 2839–2840.
- Maunakea, A.K., Nagarajan, R.P., Bilieny, M., Ballinger, T.J., D'Souza, C., Fouse, S.D., Johnson, B.E., Hong, C., Nielsen, C., Zhao, Y., Turecki, G., Delaney, A., Varhol, R., Thiessen, N., Shchors, K., Heine, V.M., Rowitch, D.H., Xing, X., Fiore, C., Schillebeek, M., Jones, S.J.M., Haussler, D., Marra, M.A., Hirst, M., Wang, T., Costello, J.F., 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* 466, 253–257.

- Mengel-From, J., Christensen, K., McGue, M., Christensen, L., 2011. Genetic variations in the *CLU* and *PICALM* genes are associated with cognitive function in the oldest old. *Neurobiology of Aging* 32, 554e7–11.
- Mercer, T.R., Dinger, M.E., Bracken, C.P., Kolle, G., Szubert, J.M., Korb, D.J., Askarian-Amiri, M.E., Gardiner, B.B., Goodall, G.J., Grimmond, S.M., Mattick, J.S., 2010. Regulated post-transcriptional RNA cleavage diversifies the eukaryotic transcriptome. *Genome Res.* 20, 1639–1650.
- Mercer, T.R., Gerhardt, D.J., Dinger, M.E., Crawford, J., Trapnell, C., Jeddell, J.A., Mattick, J.S., Rinn, J.L., 2011. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotech.* 30, 99–104.
- Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., Thomas, P.D., 2009. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.* 38, D204–D210.
- Miller, C.A., Sweatt, J.D., 2007. Covalent modification of DNA regulates memory formation. *Neuron* 53, 857–869.
- Moon, R.T., Kohn, A.D., Ferrari, G.V.D., Kaykas, A., 2004. WNT and [beta]-catenin signalling: diseases and therapies. *Nature reviews* 5, 691–701.
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNASeq. *Nat. Methods* 5, 621–628.
- Naka, H., Nakamura, S., Shimazaki, T., Okano, H., 2008. Requirement for COUP-TF and II in the temporal specification of neural stem cells in CNS development. *Nat. Neurosci.* 11, 1014–1023.
- Pal, S., Gupta, R., Kim, H., Wickramasinghe, P., Baubert, V., Showe, L.C., Dahmane, N., Davuluri, R.V., 2011. Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res.* 21, 1260–1272.
- Pareek, T.K., Belkadi, A., Kesavapany, S., Zaremba, A., Loh, S.L., Bai, L., Cohen, M.L., Meyer, C., Liby, K.T., Miller, R.H., Sporn, M.B., Letterio, J.J., 2011. Triterpenoid modulation of IL-17 and Nrf-2 expression ameliorates neuroinflammation and promotes remyelination in autoimmune encephalomyelitis. *Sci. Rep.* 201, 1–11.
- Peng, F., Yao, H., Bai, X., Zhu, X., Reiner, B.C., Beazely, M., Funa, K., Xiong, H., Buch, S., 2010. Platelet-derived growth factor-mediated induction of the synaptic plasticity gene *Arc/Arg3.1*. *J. Biol. Chem.* 285, 21615–21624.
- Pink, R.C., Wicks, K., Caley, D.P., Punch, E.K., Jacobs, L., Francisco Carter, D.R., 2011. Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA* 17, 792–798.
- Poon, S., Treweek, T.M., Wilson, M.R., Easterbrook-Smith, S.B., Carver, J.A., 2002. Clusterin is an extracellular chaperone that specifically interacts with slowly aggregating proteins on their off-folding pathway. *FEBS Lett.* 513, 259–266.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575.
- Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* 26, 841–842.
- Ramskold, D., Wang, E.T., Burge, C.B., Sandberg, R., 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS computational biology* 5, e1000598.
- Raney, B.J., Cline, M.S., Rosenbloom, K.R., Dreszer, T.R., Learned, K., Barber, G.P., Meyer, L.R., Sloan, C.A., Malladi, V.S., Roskin, K.M., Suh, B.B., Hinrichs, A.S., Clawson, H., Zweig, A.S., Kirkup, V., Fujita, P.A., Rhead, B., Smith, K.E., Pohl, A., Kuhn, R.M., Karolchik, D., Haussler, D., Kent, W.J., 2011. ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res.* 39, D871–D871.
- Rhodes, J., Lutka, F.A., Jordan-Sciutto, K.L., Bowser, R., 2003. Altered expression and distribution of *FAC1* during NGF-induced neurite outgrowth of PC12 cells. *Neuroreport* 14, 449–452.
- Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* 26, 139–140.
- Roth, R., Hevezi, P., Lee, J., Willhite, D., Lechner, S., Foster, A., Zlotnik, A., 2006. Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics* 7, 67–80.
- Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., Hume, D.A., 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature reviews* 8, 424–436.
- Schonrock, N., Matamalas, M., Ittner, L.M., Jürgen, G., 2011. MicroRNA networks surrounding APP and amyloid- $\beta$  metabolism—Implications for Alzheimer's disease. *Experimental Neurology* 235, 447–454.
- Shen, J., Bronson, R.T., Chen, D.F., Xia, W., Selkoe, D.J., Tonegawa, S., 1997. Skeletal and CNS defects in presenilin-1-deficient Mice. *Cell* 89, 629–639.
- Starkey, H., Van Kirk, C., Bidler, G., Imperio, C., Kale, V., Serfass, J., Farley, J., Yan, H., Warrington, J., Han, S., Mitschelen, M., Sonntag, W., Freeman, W., 2012. Neuroglial expression of the MHC pathway and PirB receptor is upregulated in the hippocampus with advanced aging. *Journal of Molecular Neuroscience* 48, 111–126.
- Takahashi, H., Lassmann, T., Murata, M., Carninci, P., 2012. 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protocols* 7, 542–561.
- Thomas, G.M., Huganir, R.L., 2004. MAPK cascade signalling and synaptic plasticity. *Nat. Rev. Neurosci.* 5, 173–183.
- Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., Narechania, A., 2003. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 3, 2129–2141.
- Tollervay, J.R., Curk, T., Rogelj, B., Briesse, M., Cereda, M., Kayikci, M., König, J., Hortobagyi, T., Nishimura, A.L., Zupanski, V., Patani, R., Chandran, S., Rot, G., Zupan, B., Shaw, C.E., Ule, J., 2011. Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat. Neurosci.* 14, 452–458.
- Tschan, M.P., Fischer, K.M., Fung, V.S., Pirnia, F., Borner, M.M., Fey, M.F., Tobler, A., Torbett, B.E., 2003. Alternative splicing of the human cyclin D-binding Myb-like protein (hDMP1) yields a truncated protein isoform that alters macrophage differentiation patterns. *J. Biol. Chem.* 278, 42750–42760.
- Valen, E., Pascarella, G., Chalk, A., Maeda, N., Kojima, M., Kawazu, C., Murata, M., Nishiyori, H., Lazarevic, D., Motti, D., Marstrand, T.T., Tang, M.H., Zhao, X., Krogh, A., Winther, O., Arakawa, T., Kawai, J., Wells, C., Daub, C., Harbers, M., Hayashizaki, Y., Gustincich, S., Sandelin, A., Carninci, P., 2009. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome research* 19, 255–265.
- Vaquerezas, J.M., Kummerfeld, S.K., Teichmann, S.A., Luscombe, N.M., 2009. A census of human transcription factors: function, expression and evolution. *Nature reviews* 10, 252–263.
- Yerbury, J.J., Poon, S., Meehan, S., Thompson, B., Kumita, J.R., Dobson, C.M., Wilson, M.R., 2007. The extracellular chaperone clusterin influences amyloid formation and toxicity by interacting with prefibrillar structures. *FASEB J.* 21, 2312–2322.

# Chapter 3

## A promoter level mammalian expression atlas

Published as Forrest ARR, Kawaji H, Rehli M, Baillie K, de Hoon MJL, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, Andersson R, Mungall CJ, Meehan TF, Schmeier S, Bertin N, Jørgensen M, Dimont E, Arner E, Schmidl C, Schaefer U, Medvedeva YA, Plessy C, Vitezic M, Severin J, Semple CA, Ishizu Y, Young RS, Francescato M, et al. 2014. A promoter level mammalian expression atlas. Nature 507:462–470.





# A promoter-level mammalian expression atlas

The FANTOM Consortium and the RIKEN PMI and CLST (DGT)\*

Regulated transcription controls the diversity, developmental pathways and spatial organization of the hundreds of cell types that make up a mammal. Using single-molecule cDNA sequencing, we mapped transcription start sites (TSSs) and their usage in human and mouse primary cells, cell lines and tissues to produce a comprehensive overview of mammalian gene expression across the human body. We find that few genes are truly 'housekeeping', whereas many mammalian promoters are composite entities composed of several closely separated TSSs, with independent cell-type-specific expression profiles. TSSs specific to different cell types evolve at different rates, whereas promoters of broadly expressed genes are the most conserved. Promoter-based expression analysis reveals key transcription factors defining cell states and links them to binding-site motifs. The functions of identified novel transcripts can be predicted by coexpression and sample ontology enrichment analyses. The functional annotation of the mammalian genome 5 (FANTOM5) project provides comprehensive expression profiles and functional annotation of mammalian cell-type-specific transcriptomes with wide applications in biomedical research.

The mammalian genome encodes the instructions to specify development from the zygote through gastrulation, implantation and generation of the full set of organs necessary to become an adult, to respond to environmental influences, and eventually to reproduce. Although the genome information is the same in almost all cells of an individual, at least 400 distinct cell types<sup>1</sup> have their own regulatory repertoire of active and inactive genes. Each cell type responds acutely to alterations in its environment with changes in gene expression, and interacts with other cells to generate complex activities such as movement, vision, memory and immune response.

Identities of cell types are determined by transcriptional cascades that start initially in the fertilised egg. In each cell lineage, specific sets of transcription factors are induced or repressed. These factors together provide proximal and distal regulatory inputs that are integrated at transcription start sites (TSSs) to control the transcription of target genes. Most genes have more than one TSS, and the regulatory inputs that determine TSS choice and activity are diverse and complex (reviewed in ref. 2).

Unbiased annotation of the regulation, expression and function of mammalian genes requires systematic sampling of the distinct mammalian cell types and methods that can identify the set of TSSs and transcription factors that regulate their utilization. To this end, the FANTOM5 project has performed cap analysis of gene expression (CAGE)<sup>3</sup> across 975 human and 399 mouse samples, including primary cells, tissues and cancer cell lines, using single-molecule sequencing<sup>3</sup> (Fig. 1; see the full sample list in Supplementary Table 1).

CAGE libraries were sequenced to a median depth of 4 million mapped tags per sample (Supplementary Methods) to produce a unique gene expression profile, focused specifically on promoter utilization. CAGE has advantages over RNA-seq or microarrays for this purpose, because it permits separate analysis of multiple promoters linked to the same gene<sup>13</sup>. Moreover, we show in an accompanying manuscript<sup>4</sup> that the data can be used to locate active enhancers, and to provide numerous insights into cell-type-specific transcriptional regulatory networks (see the FANTOM5 website <http://fantom.gsc.riken.jp/5>). The data extend and complement the recently published ENCODE<sup>5</sup> data, and

microarray-based gene expression atlases<sup>6</sup> to provide a major resource for functional genome annotation and for understanding the transcriptional networks underpinning mammalian cellular differentiation.

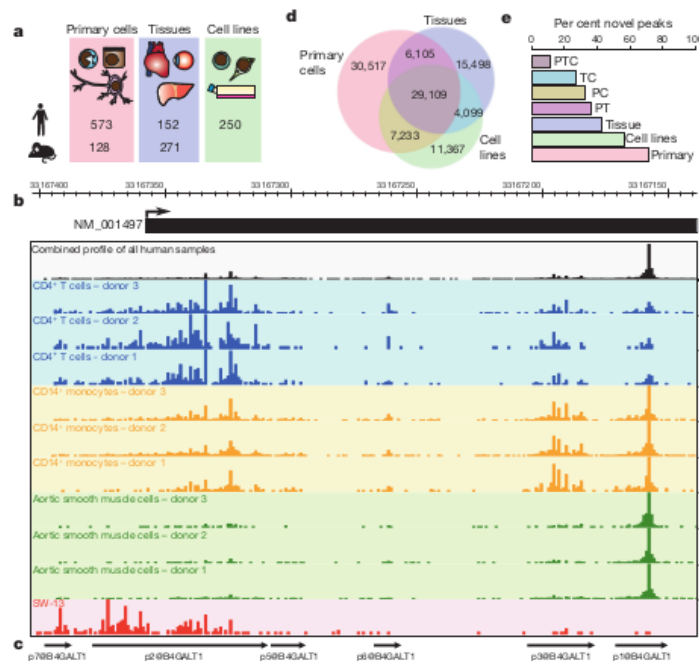
## The FANTOM5 promoter atlas

Single molecule CAGE profiles were generated across a collection of 573 human primary cell samples (~3 donors for most cell types) and 128 mouse primary cell samples, covering most mammalian cell steady states. This data set is complemented with profiles of 250 different cancer cell lines (all available through public repositories and representing 154 distinct cancer subtypes), 152 human post-mortem tissues and 271 mouse developmental tissue samples (Fig. 1a; see the full sample list in Supplementary Table 1). To facilitate data mining all samples were annotated using structured ontologies (Cell Ontology<sup>7</sup>, Uberon<sup>8</sup>, Disease Ontology<sup>9</sup>). The results of all analyses are summarized in the FANTOM5 online resource (<http://fantom.gsc.riken.jp/5>). We also developed two specialized tools for exploration of the data. ZENBU, based on the genome browser concept, allows users to interactively explore the relationship between genomic distribution of CAGE tags and expression profiles<sup>10</sup>. SSTAR, an interconnected semantic tool, allows users to explore the relationships between genes, promoters, samples, transcription factors, transcription factor binding sites and coexpressed sets of promoters. These and other ways to access the data are described in more detail in Supplementary Note 1.

## CAGE peak identification and thresholding

To identify CAGE peaks across the genome we developed decomposition-based peak identification (DPI; described in Supplementary Methods; Extended Data Fig. 1). This method first clusters CAGE tags based on proximity. For clusters wider than 49 base pairs (bp) it attempts to decompose the signal into non-overlapping sub-regions with different expression profiles using independent component analysis<sup>11</sup>. Sample- and genome-wide, DPI identified 3,492,729 peaks in human and 2,088,255 peaks in mouse. To minimize the fraction of peaks<sup>3</sup> that map to internal exons (which could exist due to post-transcriptional cleavage and recapping of RNAs<sup>12</sup>), and enrich for TSSs, we applied tag evidence thresholds

\*Lists of participants and their affiliations appear at the end of the paper.



**Figure 1 | Promoter discovery and definition in FANTOM5.** **a**, Samples profiled in FANTOM5. **b**, Reproducible cell-type-specific CAGE patterns observed for the 266 base CpG island associated *B4GALT1* locus transcription initiation region hg19:chr9:33167138..33167403. CAGE profiles for CD4<sup>+</sup> T cells (blue), CD14<sup>+</sup> monocytes (gold), aortic smooth muscle cells (green) and the adrenal cortex adenocarcinoma cell line SW-13 (red) are shown. A combined pooled profile showing TSS distribution across the entire human collection is shown in black. Values on the y axis correspond to maximum normalized TPM for a single base in each track. **c**, Decomposition-based peak identification (DPI) finds 6 differentially used peaks within this composite transcription initiation region (note: peaks are labelled from p1@B4GALT1

with most tag support through to p7@B4GALT1 with the least tag support; p4@B4GALT1 is not shown and is in the 3' UTR of the locus at position hg19:chr9:33111241..33111254–). Note in particular one large broad region on the left used in all samples and a sharp peak to the right, preferentially used in the aortic smooth muscle cells. **d**, Venn diagram showing DPI defined peaks expressed at ≥ 10 TPM in primary cells (red), tissues (blue) and cell lines (green). **e**, Fraction of unannotated peaks observed in subsets of **d**. P, primary cells; T, tissues; C, cell lines; PT, TC, PC and PTC correspond to peaks found in multiple sample types, for example, PT, found in primary cells and tissue samples.

to define robust and permissive subsets (described in more detail in Supplementary Methods and summarized in Table 1). Specifically the robust threshold, which is used for most of the analyses presented here, enriched for peaks at known 5' ends compared to known internal exons by twofold (that is, two-thirds of the peaks hitting known full-length transcript models hit the 5' end). A flow diagram showing the relationship between samples, peaks, thresholding and subsets used in each analysis is provided in the Supplementary Figure 1. Supporting evidence that the peaks are genuine TSSs, based upon support from expressed sequence tags (ESTs), histone H3 lysine 4 trimethylation (H3K4Me3) marks and DNase hypersensitive sites is provided in Supplementary Note 2.

Figure 1b illustrates the 266 bp spanning transcription initiation region of *B4GALT1*, where 6 independent robust peaks were identified by DPI, each with a unique regulatory pattern (Fig. 1c). A total of 58% of human and 56% of mouse robust peaks occur in such composite transcription initiation regions, defined as clusters of robust peaks within 100 bases of each other. More than half of these contain peaks with statistically significant differences in expression profiles (63% of human and 54% of mouse composite transcription initiation regions; likelihood ratio test, false discovery rate (FDR) < 1%, Extended Data Fig. 1d). Supplementary Tables 2 and 3 summarize public domain EST evidence that these independent peaks contained within composite transcription initiation regions give rise to long RNAs.

### Known gene coverage in FANTOM5

To provide annotation of the CAGE peaks, the distance between individual peaks and the 5' ends of known full-length transcripts was determined and then peaks within 500 bases of the 5' end of known transcript models were assigned to that gene (see Supplementary Methods, Table 1). To provide names for each TSS region, peaks identified at the permissive threshold were ranked by the total number of tags supporting each and then sequentially numbered (for example, p1@GFAP corresponds to the promoter of *GFAP* which has the highest tag support). From these annotations, TSS for 91% of human protein coding genes (as defined by the HUGO Gene Nomenclature Committee) were supported by robust CAGE peaks, and 94% at the permissive threshold (Supplementary Note 3). The atlas also detected signals from the promoters of short RNA primary transcripts, and long non-coding RNAs. In comparison to the previous FANTOM3 and 4 projects, FANTOM5 measured expression at an additional 4,721 human and 5,127 mouse RefSeq genes. The inclusion of primary cells, human and tissues in the atlas provided greater coverage than any of the sample types alone (Fig. 1d) and the primary cell samples in particular were a rich source of unannotated peaks (Fig. 1e).

### Mammalian promoter architectures

Mammalian promoters can be classified as broad or sharp types, based upon local spread of TSSs along the genome<sup>13</sup>. The FANTOM5 data

**Table 1** | Summary of peaks, coverage and genes hit in FANTOM5

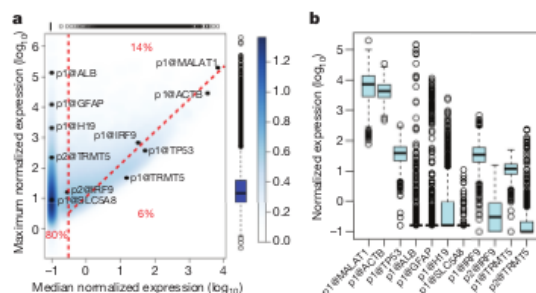
	Human						Mouse					
	Peaks	Stranded genome coverage (bp)		Number of aligned reads	Genes hit	Peaks per gene	Peaks	Stranded genome coverage (bp)		Number of aligned reads	Genes hit	Peaks per gene
The whole genome	—	$6.2 \times 10^9$	100%	$4.5 \times 10^9$	100%	—	—	$5.3 \times 10^9$	100%	$1.9 \times 10^9$	100%	—
'Permissive' CAGE peaks	1,048,124	$1.4 \times 10^7$	0.22%	$3.6 \times 10^9$	80%	20,808	652,860	$8.4 \times 10^6$	0.16%	$1.5 \times 10^9$	79%	20,480
(A) Within 500bp of annotated 5'	245,514	$4.3 \times 10^6$	0.07%	$3.0 \times 10^9$	68%	20,808	146,185	$2.5 \times 10^6$	0.05%	$1.3 \times 10^9$	69%	20,480
(B) TSS classifier positive	217,572	$4.0 \times 10^6$	0.06%	$2.9 \times 10^9$	64%	18,503	129,466	$2.4 \times 10^6$	0.05%	$1.0 \times 10^9$	52%	17,088
(A or B) Likely TSS	308,214	$5.3 \times 10^6$	0.09%	$3.2 \times 10^9$	72%	20,808	173,564	$3.0 \times 10^6$	0.06%	$1.4 \times 10^9$	70%	20,480
'Robust' CAGE peaks	184,827	$3.9 \times 10^6$	0.06%	$3.5 \times 10^9$	77%	18,961	116,277	$2.5 \times 10^6$	0.05%	$1.4 \times 10^9$	75%	19,001
(A) Within 500bp of annotated 5'	82,150	$2.2 \times 10^6$	0.04%	$3.0 \times 10^9$	66%	18,961	61,134	$1.6 \times 10^6$	0.03%	$1.3 \times 10^9$	68%	19,001
(B) TSS classifier positive	76,445	$2.1 \times 10^6$	0.03%	$2.9 \times 10^9$	63%	17,285	51,611	$1.4 \times 10^6$	0.03%	$9.9 \times 10^8$	51%	16,028
(A or B) Likely TSS	92,783	$2.4 \times 10^6$	0.04%	$3.2 \times 10^9$	70%	18,961	77,674	$1.7 \times 10^6$	0.03%	$1.3 \times 10^9$	69%	19,001
Cross-species projected robust peaks	70,351	$1.6 \times 10^6$	0.03%	—	—	—	105,157	$2.4 \times 10^6$	0.04%	—	—	—
'Homologous' robust peaks	34,041	$1.0 \times 10^6$	0.02%	—	—	—	42,423	$1.3 \times 10^6$	0.02%	—	—	—

confirmed this general observation (Extended Data Fig. 2), however, for the first time the greater depth of sequencing enabled identification of the preferred TSS within broad promoters. Taking each library in turn, using the location of the dominant TSS (that is, the TSS with the highest number of tags), we searched for phased WW dinucleotides (AA/AT/TA/TT) associated with nucleosome location<sup>14</sup> (Extended Data Fig. 2). Remarkably, on a genome-wide scale, there was a periodic spacing of WW motifs with a 10.5 bp repeat downstream of the dominant TSS, exactly as shown previously for well-phased H2A.Z nucleosomes<sup>14</sup> (Extended Data Fig. 2d). The precise phasing was supported further by the pattern of H2A.Z and H3K4me3 chromatin immunoprecipitation sequencing (ChIP-seq) signal seen around TSS in CD14<sup>+</sup> monocytes and frontal lobe respectively (Extended Data Fig. 2e, f). This observation indicates that the positioned nucleosome is a key indicator of start site preference in broad promoters.

### Expression levels and tissue specificity

The raw tag counts under the DPI peak coordinates were used to generate an expression table across the entire collection. Normalized tags per million (TPM) were then calculated using the relative log expression (RLE) method in edgeR<sup>15</sup>. Almost all peaks (96%) were reproducibly detected above 1 TPM in at least two samples, but most were detected in less than half the samples. Examining the distribution of expression level and breadth across the collection, we classified the 185K robust human peak expression profiles as non-ubiquitous (cell-type-restricted, 80%), ubiquitous-uniform ('housekeeping', 6%) or ubiquitous-non-uniform (14%) (Fig. 2a, b). We define ubiquitous as detected in more than 50% of samples (median >0.2 TPM) and uniform as a less than tenfold difference between maximum and median expression. Estimation using the smaller mouse expression data set or human primary cell, cell line or tissue data subsets resulted in different fractions, yet in all cases ubiquitous-uniform expression profiles were in the minority (Extended Data Fig. 3a–e). Alternative measures such as richness index and Shannon entropy confirm that only a minor fraction of transcripts can be considered as genuine housekeeping genes with broad and uniform expression (Supplementary Note 4 and Supplementary Table 4 for a

list of housekeeping genes). In addition many of the 1,225 known genes that were missed in the collection are known to be specifically expressed in cell types that are not easily procured; indicating that even more of the mammalian transcriptome has a cell-type-restricted expression





pattern (Supplementary Note 3). In overview, the data confirm the argument that most genes are regulated in a tissue-dependent manner<sup>16</sup>. According to Gene Ontology enrichment analysis<sup>17</sup> of genes within each of the three classes (Supplementary Table 5), the non-ubiquitous genes were enriched for proteins involved in cell–cell signalling, plasma membrane receptors, cell adhesion molecules and signal transduction, whereas genes in the housekeeping set were enriched for components of the ribonucleoprotein complex and RNA processing. The ubiquitous-non-uniform set was enriched for cell cycle genes, with 204 of the 268 human genes annotated with the 'mitotic cell cycle' term, a reflection of the fact that the fraction of actively proliferating cells inevitably varies greatly across the collection.

Finally, of the 104,859 peaks expressed at 10 TPM (~3 copies per cell<sup>18</sup>) or greater, an average primary cell sample expressed a median of 8,757 including peaks for 430 transcription factor mRNAs (Extended Data Fig. 3f, g).

### Promoter conservation between human and mouse

Regulatory regions such as transcription factor binding sites are often, but not always, located in conserved and orthologous regions<sup>19</sup>. Overall human TSSs were significantly enriched in evolutionarily conserved regions compared to the genome-wide null expectation, with 38% overlapping previously defined mammalian constrained elements (Fisher's exact test, odds ratio 10.2,  $P$  value  $< 2.2 \times 10^{-16}$ ; see Supplementary Methods). Despite this general level of conservation, there is evidence of extensive evolutionary remodelling of transcription initiation. For example, 43% (79,670 out of 184,476) of human TSSs could not be aligned to the mouse genome, and 39% (45,926 out of 116,277) of mouse TSSs could not be aligned to the human genome (Supplementary Methods). Alignment between species decayed as a function of neutral sequence divergence (Fig. 3). Housekeeping TSSs showed highest TSS conservation, whereas the TSSs of non-coding RNAs were less conserved than those of protein-coding TSSs. Indeed, the alignment of promoters of

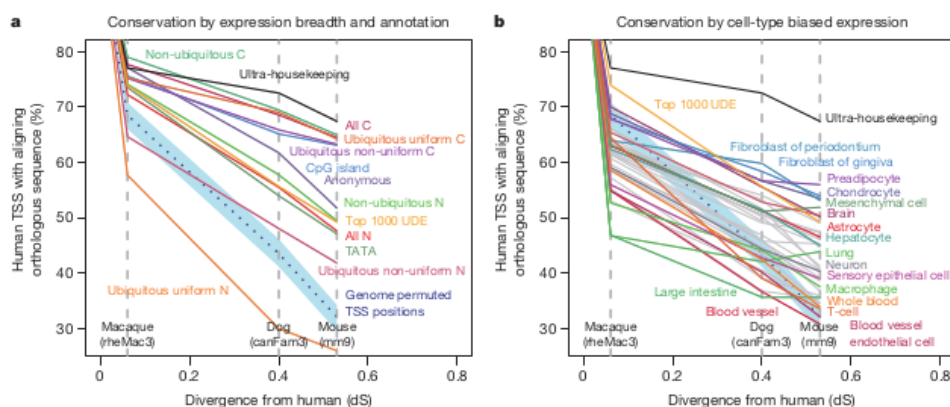
broadly expressed non-coding transcripts was not greatly different from randomly selected genomic sites (Fig. 3a). However, it is important to note that the random permutations inevitably overlap constrained elements, so cannot be considered representative of neutral evolution.

TSSs that were highly-restricted or biased in their expression to a single cell type or tissue were more likely to be gained or lost through evolution (Fig. 3a). TSSs preferentially expressed in fibroblasts, chondrocytes and pre-adipocytes were among the most conserved, whereas those enriched in T-cells, macrophages, dendritic cells, whole blood and endothelial cells were the most likely to be gained or lost (Fig. 3b). This suggests a more rapidly evolving immune system. It also suggests contributions of relaxed constraint and positive selection to the remodelling of transcription initiation through the insertion and deletion of promoter sequences.

To enable comparative analysis, we projected the expression patterns from one species to the other (Extended Data Fig. 4) and provide the peak position and orthologous expression profile through a cross-species track in ZENBU<sup>10</sup>. Only 54% and 61% of human and mouse conserved TSSs (of protein coding genes) had an orthologous peak in the other species. This increased to 61% and 63% respectively for TSSs from well matched samples (for example, human and mouse hepatocytes), however, surprisingly, almost 40% of conserved TSS do not appear to be used even in the matched cells (Supplementary Table 6).

### Features of cell-type-specific promoters

Carrying out a systematic *de novo* motif discovery analysis in cell-type-specific promoters, recovered motifs similar to the binding motifs of transcription factors known to be relevant to the corresponding cellular states (Extended Data Fig. 5a–c and described in Supplementary Note 5). Examining general promoter features many CpG island (CGI) based promoters (54%) and most non-CGI-non-TATA promoters (92%) had non-ubiquitous expression profiles (Extended Data Fig. 3k–n). Although CGI promoters are generally associated with housekeeping



**Figure 3 | TSS conservation as a function of expression properties and functional annotation.** **a, b**, Human robust TSS coordinates were projected through EPO12 whole genome multiple sequence alignments (Supplementary Methods). The y-axis values show the fraction of human TSSs that align to an orthologous position in the indicated species. The x axis shows the relative divergence of macaque, dog and mouse genomes as the substitution rate at fourfold degenerate sites in protein coding sequence. The TSS locations were genome permuted (Supplementary Methods) and then projected through EPO12 alignments to give the null expectation (dashed blue line). The 95% confidence intervals of 1,000 samples of 1,000 TSS are shown (blue shading). **a**, TSS mapped to the 5' ends of protein coding and non-coding transcripts are labelled (C and N, respectively), those that do not map to a known transcript 5' end are shown as the 'anonymous' category. With the exception of

anonymous, all robust TSSs represented in both panels are associated with the 5' ends of previously annotated transcripts. Non-ubiquitous (cell-type-restricted), ubiquitous-uniform (housekeeping) and non-uniform-ubiquitous were defined as in Fig. 2. Ultra-housekeeping TSSs were defined as those with less than fivefold difference between maximum and median. The category top 1000 UDE represents the 1,000 ubiquitous TSSs that are most differentially expressed. There are 1,016 ultra-housekeeping TSSs, 276 ubiquitous-uniform non-coding TSSs and all other categories contain over 2,000 TSSs. **b**, Same axes as panel **a** showing TSSs with expression that is biased towards a single expression facet (larger mutually exclusive grouping of the primary cell and tissue samples based on the sample ontologies CO and UBERON, defined in ref. 4). Only expression facets with greater than 250 enriched TSSs are shown. For clarity, only a subset of expression facets are coloured and labelled.

genes, we observed a subset with highly cell-type-restricted expression profiles (right tail of Extended Data Fig. 6a). Examining CGI and non-CGI promoters separately we find that cell-type-specific promoters of both classes were enriched for binding of cell-type-specific transcription factors (evidenced by over-representation of motifs and bound sites in public ChIP-seq data sets). For the human hepatocellular carcinoma cell line HepG2 we observed enrichment of liver-specific transcription factors (HNF4, FOXA2, and TCF7L2) at both CGI and non-CGI HepG2 specific promoters (Extended Data Fig. 6b, c; similar examples are shown in Extended Data Figs 5d and 7). As noted in the accompanying analysis<sup>4</sup>, both cell-type-specific CGI and non-CGI promoters tend to have proximal high-specificity enhancers (Extended Data Fig. 6d). This indicates that specific expression at CGI promoters uses the same type of signals as non-CGI promoters: proximal transcription factor motifs and high-specificity enhancers.

Of note, a small number of highly abundant RNAs account for 20% or more of the reads in some libraries: HBB, SMR3B, STATH, PRB4, CLPS, HTN3, SERPINA1, CTRB2, CPB1, CPA1 and MALAT1. Although the abundance of these transcripts is a function of their relative stability as well as rate of initiation, a modest but significant over representation of ETS and YY1 sites was found in highly expressed promoters compared to weakly expressed ones (Extended Data Fig. 5g). Although the different motif composition may contribute to expression levels, the accompanying manuscript<sup>4</sup> shows that arrays of enhancers with similar usage<sup>20</sup> probably contribute to the higher maximal expression rate.

### Key cell-type-specific transcription factors

Among 1,762 human and 1,516 mouse transcription factors compiled from the literature<sup>21–23</sup>, promoter level expression profiles for 1,665 human transcription factors (94%) and 1,382 mouse transcription factors (91%) were obtained (Supplementary Tables 7, 8 and 9 and Supplementary Note 6). The distribution of expression levels and cell-type or tissue-specificity of transcription factors (Extended Data Fig. 3f–j) and the number of robust promoter peaks per transcription factor gene was similar to coding genes in general (4.8 compared to 4.6). In any given primary cell type, a median of 430 (306 to 722) transcription factors were expressed at 10 TPM or above (~3 copies per cell based on 300,000 mRNAs per cell<sup>18</sup>) (Extended Data Fig. 3g).

Clustering transcription factors by expression profile revealed sets of transcription factors specifically enriched in each cell type (Extended Data Fig. 8). For each primary cell sample we have made available ranked lists of transcription factors based on their promoter expression in the sample relative to the median across the collection ([http://fantom.gsc.riken.jp/5/ssstar/Browse\\_samples](http://fantom.gsc.riken.jp/5/ssstar/Browse_samples)). For most cell types we found one transcription factor that was very highly enriched ( $\geq 100$ -fold), 23 highly enriched transcription factors ( $\geq 10$ -fold) and 82 moderately enriched transcription factors ( $\geq 5$ -fold) (numbers of transcription factors are based on median number of transcription factors observed at each enrichment threshold across the primary cell samples). To demonstrate their likely relevance we systematically reviewed phenotypes of transcription factor knockout mice at the MGI (see Supplementary Note 7). The clear connection between tissue-specific expression profiles and relevant knockout phenotypes is summarized in Supplementary Table 10. For example, in mouse inner ear hair cells, knockout of six of the top 20 most enriched transcription factor genes in mouse (*Pou3f4* (ref. 24), *Sox2* (ref. 25), *Egr2*, *Six1* (ref. 26), *Fos*<sup>27</sup>, *Tbx18* (ref. 28)) as well as patient mutations in a further four top transcription factor genes (*POU4F3* (ref. 29), *ZIC2* (ref. 30), *SOX10* (ref. 31), *FOXF2* (ref. 32)) resulted in hearing-related defects. Similarly, mouse knockouts or patients with mutations in the transcription factors enriched in osteoblasts (*CREB3L1* (ref. 33), *DLX5* (ref. 34), *EBF2* (ref. 35), *HAND2* (ref. 36), *HOXC5* (ref. 37), *NFIX*<sup>38</sup>, *PRRX1* (ref. 39), *PRRX2* (ref. 40), *SIX1* (ref. 41), *TWIST1* (ref. 42), *SHOX*<sup>43</sup>, *Six2* (ref. 44)) had bone and osteoblast phenotypes. A substantial fraction of top transcription factors (61% of mouse and 40% of human transcription factors) have relevant phenotypes recorded in knockout mice (Supplementary Table 10).

### Inferring function from expression profiles

Taking a pair-wise Pearson correlation matrix of the promoter expression profiles we carried out MCL clustering<sup>45</sup> (Supplementary Methods) to group promoters that share similar expression profiles across the atlas. Figure 4 shows a graphical overview of the structure of the data (and the mouse counterpart is shown in Extended Data Fig. 9). We find 6,030 cases of named genes with alternative promoters participating in two or more coexpression clusters (Extended Data Fig. 10). To evaluate and annotate these coexpressed groups, we tested for enrichment in specific Gene Ontology terms and in a curated database of 489 biological pathways. Of these, 356 pathways (174 KEGG, 114 WikiPathways, 46 Reactome, 22 Netpath) were significantly enriched in at least one human coexpression group (FDR < 0.05). Using this approach, 38% of the unannotated robust peaks (35,082 out of 91,269) were within a cluster with a significant association to a pathway. The annotated coexpression groups are summarized in the website ([http://fantom.gsc.riken.jp/5/ssstar/Browse\\_coexpression\\_clusters](http://fantom.gsc.riken.jp/5/ssstar/Browse_coexpression_clusters)) and a detailed example identifying genes putatively involved in influenza A pathogenesis is shown in Extended Data Fig. 10a.

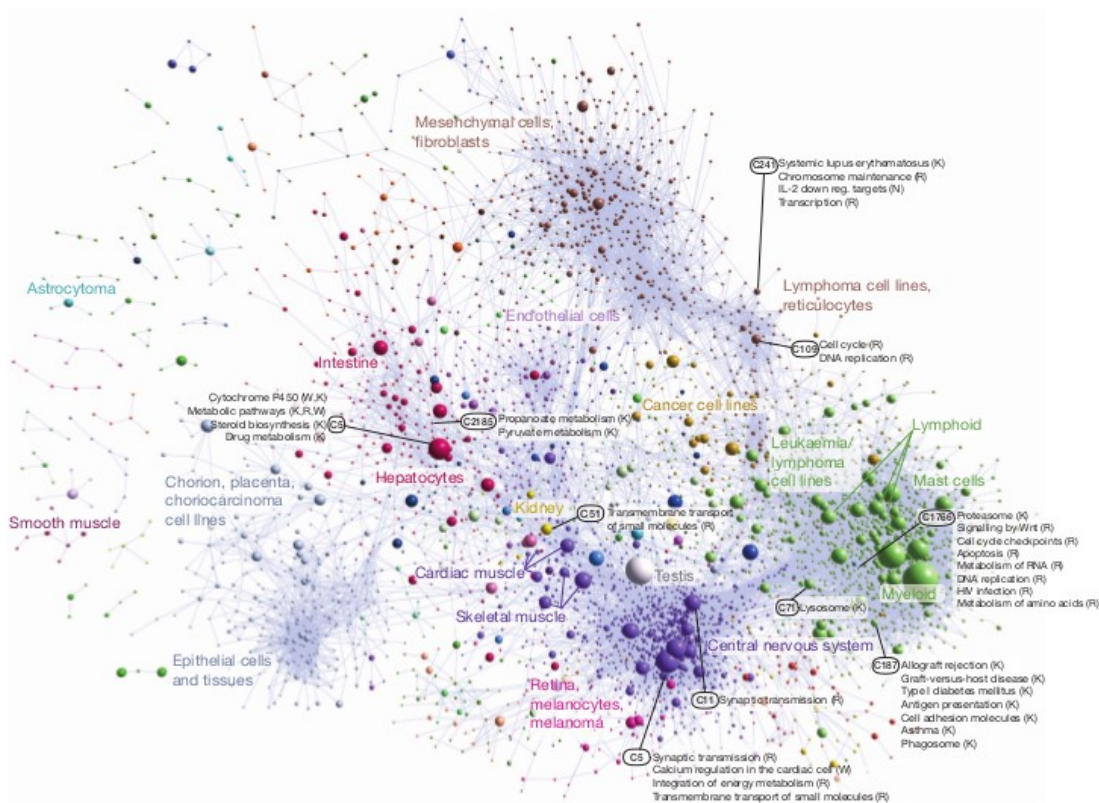
Introducing sample ontology enrichment analysis (SOEA), we show that expression profiles can also be associated with cell, anatomical and disease ontology terms by testing for overrepresentation of terms in ranked lists of systematically annotated samples expressing each peak (Extended Data Fig. 11 and Supplementary Methods). Novel peaks can be annotated in this way. For example, an un-annotated DPI peak at hg19:chr18:3659943..3659972, + is linked to the terms classical monocyte (CL:0000860;  $P$  value =  $6.35 \times 10^{-124}$ , Extended Data Fig. 11h) and bone marrow (UBERON:0002371;  $P$  value =  $2.7 \times 10^{-80}$ ). Manual examination of the profile confirms the transcript is predominantly expressed in myeloid cells with higher levels in CD14<sup>+</sup> monocytes. Applied to all CAGE peaks, 127,645 human and 44,449 mouse robust peaks were annotated as enriched in at least one CL, DOID or UBERON term (Extended Data Fig. 11i, j). The most commonly-enriched terms at a  $P$  value threshold of  $10^{-20}$  were classical monocyte (CL:0000860; 26,634 peaks, 14%), bone marrow (UBERON:0002371; 22,387 peaks, 12%) and neural tube (UBERON:0001049; 20,484 peaks, 11%) (Supplementary Table 13). This is consistent with the coexpression clustering in Fig. 4 (green and purple spheres correspond to leukocyte and central nervous system enriched expression profiles) and indicates that a large fraction of the mammalian genome is dedicated to immune and nervous system specific functions.

### Conclusion

The FANTOM5 promoter atlas is a natural extension of earlier maps of active transcripts and promoters complementing the sequencing of mammalian genomes<sup>46,47</sup>. It represents an advance in an order of magnitude in the wide range of cell types and the amount of data produced per sample, and using single-molecule sequencing avoided polymerase chain reaction (PCR), digestion and cloning bias<sup>48</sup>. We have identified and quantified the activity of at least one promoter for more than 95% of annotated protein-coding genes in the human reference genome; only the activity of 1,225 promoters remains uncharacterized. Some of these may not actually be expressed. Some cannot be unambiguously measured with CAGE due to copy number variants or closely related multigene families. The remaining promoters are probably expressed in rare cell types or during windows of development or states of cellular activation that are not readily accessible and remain to be sampled. A continued effort to add profiles from these cells will make it possible to integrate them with the FANTOM5 data, and to extract metadata to identify those regulatory elements that are new and lineage-specific.

The FANTOM5 data highlights the value in profiling primary cells as opposed to whole tissues. It also highlights the weakness of using cancer cell lines. The cancer cell lines generally fail to cluster in a sample-to-sample correlation graph with their supposed cell type or tissue of origin (Extended Data Fig. 12) and express more transcription factors than primary cells (Extended Data Fig. 3g). The mutations and





**Figure 4 | Coexpression clustering of human promoters in FANTOM5.** Collapsed coexpression network derived from 4,882 coexpression groups (one node is one group of promoters; 4,664 groups are shown here) derived from expression profiles of 124,090 promoters across all primary cell types, tissues and cell lines (visualized using Biocluster Express<sup>45</sup> (ref. 45),  $r > 0.75$ , MCLi = 2.2). For display, each group of promoters is collapsed into a sphere, the radius of which is proportional to the cube root of the number of promoters

in that group. Edges indicate  $r > 0.6$  between the average expression profiles of each cluster. Colours indicate loosely-associated collections of coexpression groups (MCLi = 1.2). Labels show representative descriptions of the dominant cell type in coexpression groups in each region of the network, and a selection of highly-enriched pathways ( $FDR < 10^{-4}$ ) from KEGG (K), WikiPathways (W), Netpath (N) and Reactome (R). Promoters and genes in the coexpression groups are available online at (<http://fantom.gsc.riken.jp/5/data/>).

chromosomal rearrangements that occur in cancer result in unique transcriptional networks that do not exist in the untransformed state and do not necessarily generalize across multiple tumours of the same type. In terms of building mammalian transcriptional regulatory network models that reflect the normal untransformed state, primary cells are the logical choice. They have normal genomes, and express in the order of 430 transcription factors at appreciable levels, ranking of which can be used to reduce the complexity further and identify key known regulators of cellular phenotypes. Focusing on these key regulators and motif searching in the corresponding cell-type-specific promoters provides the data to build cell-type-specific regulatory network models and support a rational approach to identification of drivers required to reprogram cells from one lineage to another. Promoter-based expression data also has direct practical applications in the interpretation (and re-interpretation) of the function of single nucleotide polymorphisms (SNPs) in genome-wide association studies (GWAS), which commonly occur in non-coding sequences. In accompanying manuscripts, reanalysis of several GWAS data sets uncovered new disease associations in FANTOM5 promoters and identification of regulatory SNPs within enhancers that were active in medically relevant samples (ref. 4 and manuscript in preparation). Accordingly, the data will enable the design of

genotyping arrays and sequence-capture systems to target regulatory variation, and the design of promoter constructs allowing researchers to specify the cell-type-specificity and absolute expression levels of their constructs (particularly for Cre-conditional knockouts<sup>49</sup> and gene therapy vectors<sup>50</sup>). In all these respects, the FANTOM5 data set greatly extends the data generated by ENCODE<sup>5</sup> to further our knowledge of genome function.

## METHODS SUMMARY

All Methods are described in full in the Supplementary Information.

**Online Content** Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 4 January 2013; accepted 26 February 2014.

- Vickaryous, M. K. & Hall, B. K. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol. Rev. Camb. Philos. Soc.* **81**, 425–455 (2006).
- Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Rev. Genet.* **13**, 233–245 (2012).
- Kanamori-Katayama, M. et al. Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res.* **21**, 1150–1159 (2011).

27 MARCH 2014 | VOL 507 | NATURE | 467

©2014 Macmillan Publishers Limited. All rights reserved

4. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* <http://dx.doi.org/10.1038/nature12787> (this issue).
5. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
6. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcripts. *Proc. Natl Acad. Sci. USA* **101**, 6062–6067 (2004).
7. Meehan, T. F. *et al.* Logical development of the cell ontology. *BMC Bioinformatics* **12**, 6 (2011).
8. Mungall, C. J., Torraia, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* **13**, R5 (2012).
9. Osborne, J. D. *et al.* Annotating the human genome with Disease Ontology. *BMC Genomics* **10** (Suppl 1), S6 (2009).
10. Severin, J. *et al.* Interactive visualization and analysis of large-scale NGS data-sets using ZENBU. *Nature Biotechnol.* <http://dx.doi.org/10.1038/nbt2840> (2014).
11. Oja, E., Hyvärinen, A. & Karhunen, J. *Independent Component Analysis* (John Wiley & Sons, 2001).
12. Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009).
13. Carrinci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* **38**, 626–635 (2006).
14. Ioshikhes, I., Hosid, S. & Pugh, B. F. Variety of genomic DNA patterns for nucleosome positioning. *Genome Res.* **21**, 1863–1871 (2011).
15. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
16. Schug, J. *et al.* Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.* **6**, R33 (2005).
17. Beissbarth, T. & Speed, T. P. GStat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20**, 1464–1465 (2004).
18. Velculescu, V. E. *et al.* Analysis of human transcriptomes. *Nature Genet.* **23**, 387–388 (1999).
19. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036–1040 (2010).
20. Barolo, S. Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy. *Bioessays* **34**, 135–141 (2012).
21. Roach, J. C. *et al.* Transcription factor expression in lipopolysaccharide-activated peripheral-blood-derived mononuclear cells. *Proc. Natl Acad. Sci. USA* **104**, 16245–16250 (2007).
22. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature Rev. Genet.* **10**, 252–263 (2009).
23. Wingender, E., Schoepf, T. & Dönitz, J. TFCClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.* **41**, D165–D170 (2013).
24. de Kok, Y. J. *et al.* Association between X-linked mixed deafness and mutations in the POU domain gene *POU3F4*. *Science* **267**, 685–688 (1995).
25. Kiernan, A. E. *et al.* *Sox2* is required for sensory organ development in the mammalian inner ear. *Nature* **434**, 1031–1035 (2005).
26. Zheng, W. *et al.* The role of *Sox1* in mammalian auditory system development. *Development* **130**, 3989–4000 (2003).
27. Paylor, R., Johnson, R. S., Papaioannou, V., Spiegelman, B. M. & Wehner, J. M. Behavioral assessment of *c-fos* mutant mice. *Brain Res.* **651**, 275–282 (1994).
28. Trowe, M. O., Maier, H., Schweizer, M. & Kispert, A. Deafness in mice lacking the T-box transcription factor *Tbx18* in otic fibrocytes. *Development* **135**, 1725–1734 (2008).
29. Vahava, O. *et al.* Mutation in transcription factor *POU4F3* associated with inherited progressive hearing loss in humans. *Science* **279**, 1950–1954 (1998).
30. Chabchoub, E., Willekens, D., Vermeesch, J. R. & Fryns, J. P. Holoprosencephaly and *ZIC2* microdeletions: novel clinical and epidemiological specificities delineated. *Clin. Genet.* **81**, 584–589 (2012).
31. Pingault, V. *et al.* *SOX10* mutations in patients with Waardenburg-Hirschsprung disease. *Nature Genet.* **18**, 171–173 (1998).
32. Kapoor, S., Mukherjee, S. B., Shroff, D. & Arora, R. Dismyelinination of the cerebral white matter with microdeletion at 6p25. *Indian Pediatr.* **48**, 727–729 (2011).
33. Murakami, T. *et al.* Signalling mediated by the endoplasmic reticulum stress transducer OASIS is involved in bone formation. *Nature Cell Biol.* **11**, 1205–1211 (2009).
34. Acampora, D. *et al.* Craniofacial, vestibular and bone defects in mice lacking the *Distal-less*-related gene *Dlx5*. *Development* **126**, 3795–3809 (1999).
35. Kieslinger, M. *et al.* *EBF2* regulates osteoblast-dependent differentiation of osteoclasts. *Dev. Cell* **9**, 757–767 (2005).
36. Funato, N. *et al.* Hand2 controls osteoblast differentiation in the branchial arch by inhibiting DNA binding of Runx2. *Development* **136**, 615–625 (2009).
37. McIntyre, D. C. *et al.* Hox patterning of the vertebrate rib cage. *Development* **134**, 2991–2999 (2007).
38. Driller, K. *et al.* Nuclear factor I $\chi$  deficiency causes brain malformation and severe skeletal defects. *Mol. Cell Biol.* **27**, 3855–3867 (2007).
39. Lu, M. F. *et al.* *prx-1* functions cooperatively with another paired-related homeobox gene, *prx-2*, to maintain cell fates within the craniofacial mesenchyme. *Development* **126**, 495–504 (1999).
40. Ten Berge, D., Brouwer, A., Korving, J., Martin, J. F. & Meijlink, F. *Px1* and *Px2* in skeletogenesis: roles in the craniofacial region, inner ear and limbs. *Development* **125**, 3831–3842 (1998).
41. Laclet, C. *et al.* Altered myogenesis in *Sox1*-deficient mice. *Development* **130**, 2239–2252 (2003).
42. Lee, M. S., Lowe, G. N., Strong, D. D., Wergedal, J. E. & Glackin, C. A. *Twist*, a basic helix-loop-helix transcription factor, can regulate the human osteogenic lineage. *J. Cell. Biochem.* **75**, 566–577 (1999).
43. Clement-Jones, M. *et al.* The short stature homeobox gene *SHOX* is involved in skeletal abnormalities in Turner syndrome. *Hum. Mol. Genet.* **9**, 695–702 (2000).
44. He, G. *et al.* Inactivation of *Sox2* in mouse identifies a novel genetic mechanism controlling development and growth of the cranial base. *Dev. Biol.* **344**, 720–730 (2010).
45. Freeman, T. C. *et al.* Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput. Biol.* **3**, e206 (2007).
46. The FANTOM Consortium. The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
47. Suzuki, H. *et al.* The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature Genet.* **41**, 553–562 (2009).
48. Kawaji, H. *et al.* Comparison of CAGE and RNA-seq transcriptome profiling using a clonally amplified and single molecule next generation sequencing. *Genome Res.* <http://dx.doi.org/10.1101/gr.156232.113> (2014).
49. Heffner, C. S. *et al.* Supporting conditional mouse mutagenesis with a comprehensive cre characterization resource. *Nature Commun.* **3**, 1218 (2012).
50. Pringle, L. A. *et al.* Rapid identification of novel functional promoters for gene therapy. *J. Mol. Med.* **90**, 1487–1496 (2012).

Supplementary Information is available in the online version of the paper.

**Acknowledgements** FANTOM5 was made possible by a Research Grant for RIKEN Omics Science Center from MEXT to Y. Hayashizaki and a grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan to Y. Hayashizaki. It was also supported by Research Grants for RIKEN Preventive Medicine and Diagnosis Innovation Program (RIKEN PMI) to Y. Hayashizaki and RIKEN Centre for Life Science Technologies, Division of Genomic Technologies (RIKEN CLST (DGT)) from the MEXT, Japan. Extended acknowledgements are provided in the Supplementary Information.

**Author Contributions** The core members of FANTOM5 phase 1 were Alistair R. R. Forrest, Hideya Kawaji, Michael Rehli, J. Kenneth Bailly, Michiel J. L. de Hoon, Timo Lassmann, Masayoshi Itoh, Kim M. Summers, Harukazu Suzuki, Carsten O. Daub, Jun Kawai, Peter Heutink, Winston Hide, Tom C. Freeman, Boris Lenhard, Vladimir B. Bajic, Martin S. Taylor, Vsevolod J. Makeev, Albin Sandelin, David A. Hume, Piero Carninci and Yoshihide Hayashizaki. Samples were provided by: A. Blumenthal, A. Bonetti, A. Mackay-sim, A. Sajantila, A. Saxena, A. Schwegmann, A.G.B., A.J.K., A.L., A.R.R.F., A.S.B.E., B.B., C. Schmidt, C. Schneider, C.A.D., C.A.W., C.K., C.L.M., D.A.H., D.A.O., D.G., D.S., D.V., E.W., F.B.N., G.G.S., G.J.F., G.S., H. Kawamoto, H. Kosaki, H. Morikawa, H. Motohashi, H. Ohno, H. Sato, H. Satoh, H. Tanaka, H. Tatsukawa, H. Toyoda, H. C.C., H.E., J. Kere, J.B., J.F., J.K.B., J.S.K., J.T., J.W.S., K.E., K.J.H., K.M., K.M.S., L.F., L.M.K., L.M.vdB., L.N.W., M. Edinger, M. Endoh, M. Fagioli, M. Hamaguchi, M. Hara, M. Herlyn, M. Morimoto, M. Rehli, M. Yamamoto, M. Yoneda, M.B., M.C.F.C., M.D., M.E.F., M.O., M.O.H., M.P., M.vdW., N.M., N.O., N.T., P.A., P.G.Z., P.H., P.R., R.F., R.G., R.K.S., R.P., R.V., S. Guhl, S. Gustincich, S. Kojima, S. Koyasu, S. Krampitz, S. Sakaguchi, S. Savvi, S.E.Z., S.O., S.P.B., S.P.K., S. Roy, S.Z., T. Kitamura, T. Nakamura, T. Nozaki, T. Sugiyama, T.B.G., T.D., T.G., T.I., T.J.H., T.J.K., V.O., W.L., Y. Hasegawa, Y. Nakachi, Y. Nakamura, Y. Yamaguchi, Y. Yonekura, Y.L., Y.L.K., Y.M. and Y.O. Analyses were carried out by: A. Mathelier, A. Meynert, A. Sandelin, A.C., A.D.D., A.P.G., A.H., A.J., A.M.B., A.P., A.R.R.F., A.S.K., A.T.K., A.V.F., B. Lenhard, B. Lilje, B.D., B.K., B.M., B.R.J., C. Schmidt, C. Schneider, C.A.S., C.F., C.J.M., C.O.D., C.P., C.V.C., D.A., D.A.M., D.C., E. Dalla, E. Dimont, E.A., E.A.S., E.J.W., E.M., E.V., E.v.N., F.D., G.J., G.J.F., G.M.A., H. Kawaji, H. Ohmiya, H. Shimoi, H.F., H.J., H.P., I.A., I.E.V., I.H., I.V.K., J.A.B., J.A.C.A., J.A.R., J.C.M., J.F.J.L., J.G., J.G.D.P., J.H., J.K.B., J.S., K. Kajiyama, K.L., K.L., L.H., L.L., M. Francescato, M. Rashid, M. Rehli, M. Roncador, M. Thompson, M.B.R., M.C., M.C.F., M.J., M.J.L.d.H., M.L., M.S.T., M.V., N.B., O.J.L.R., O.M.H., P.A.C.H., P.J.B., R.A., R.S.Y., S. Katayama, S. Kawaguchi, S. Schmeier, S. Rennie, S.F., S.J.H.S., S.P., T. Sengstag, T.C.F., T.F.M., T.H., T.K., T.L., T.R., T.T., U.S., V.B.B., V.H., W.J.M., W.H., W.W.W., X.Z., Y. Chen, Y. Ciani, Y.A.M., Y.S., Z.T. Libraries were generated by: A. Kaiho, A. Kubosaki, A. Saka, C. Simon, E.S., F.H., H.N., J. Kawai, K. Kaida, K.N., M. Furuno, M. Murata, M. Sakai, M. Tagami, M.I., M.K., M.K.K., N.K., N.N., N.S., P.C., R.M., S. Kato, S.N., S.N.-S., S.W., S.Y., T.A., T. Kawashima. The manuscript was written by A.R.R.F. and D.A.H. with help from A. Sandelin, J.K.B., M. Rehli, H.K., M.J.L.d.H., V.H., I.V.K., M.T. and K.M.S. with contributions, edits and comments from all authors. The project was managed by Y. Hayashizaki, A.R.R.F., P.C., M.I., M.S., J. Kawai, C.O.D., H. Suzuki, T.L. and N.K. The scientific coordinator was A.R.R.F. and the general organizer was Y. Hayashizaki.

**Author Information** All CAGE data has been deposited at DDBJ DRA under accession number DRA000991. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.R.R.F. (alistair.forrest@gmail.com), P.C. (carninci@riken.jp) or Y.H. (yoshihide@gscl.riken.jp).

**The FANTOM Consortium and the RIKEN PMI and CLST (DGT)**

Alistair R. R. Forrest<sup>1,2\*</sup>, Hideya Kawaji<sup>1,2,3\*</sup>, Michael Rehli<sup>4,5\*</sup>, J. Kenneth Bailly<sup>6\*</sup>, Michiel J. L. de Hoon<sup>1,2</sup>, Vanja Haberle<sup>7,8</sup>, Timo Lassmann<sup>1,2</sup>, Ivan V. Kulakovskiy<sup>9,10</sup>, Marina Lizio<sup>1,2</sup>, Masayoshi Itoh<sup>1,2,3</sup>, Robin Andersson<sup>1,11</sup>, Christopher J. Mungall<sup>12</sup>, Terrence F. Meehan<sup>13</sup>, Sebastian Schmeier<sup>14,15</sup>, Nicolas Bert<sup>1,2</sup>, Mette Jørgensen<sup>11</sup>, Emmanuel Dimont<sup>16</sup>, Erik Amer<sup>1,2</sup>, Christian Schmidt<sup>17</sup>, Ulf Schaefer<sup>1,4</sup>, Yulia A. Medvedeva<sup>10,14</sup>, Charles Plessy<sup>1,2</sup>, Morana Vitezic<sup>1,17</sup>, Jessica Severin<sup>1,2</sup>, Colin A. Semple<sup>18</sup>, Yuri Ishizu<sup>1,2</sup>, Robert S. Young<sup>18</sup>, Margherita Francescato<sup>19,20</sup>, Intikhab Alam<sup>1,4</sup>, Davide Albanese<sup>21</sup>, Gabriel M. Altschuler<sup>19</sup>, Takahiro Arakawa<sup>1,2</sup>, John A. C.



- Archer<sup>1,4</sup>, Peter Arner<sup>2,22</sup>, Magda Babina<sup>2,3</sup>, Sarah Rennie<sup>1,8</sup>, Piotr J. Balwiercz<sup>2,4</sup>, Anthony G. Beckhouse<sup>2,5,26</sup>, Swati Pradhan-Bhatt<sup>2,7</sup>, Judith A. Blake<sup>2,8</sup>, Antje Blumenthal<sup>26,29</sup>, Beatrice Bodeg<sup>30</sup>, Alessandro Bonetti<sup>1,2</sup>, James Briggs<sup>25,†</sup>, Frank Brombacher<sup>31,32</sup>, A. Maxwell Burroughs<sup>1</sup>, Andrea Califano<sup>33,34,35,36</sup>, Carlo V. Cannistraci<sup>3,7,38,†</sup>, Daniel Carballo<sup>39</sup>, Yun Chen<sup>41</sup>, Marco Chierici<sup>2,1</sup>, Yan Ciani<sup>40</sup>, Hans C. Clevers<sup>41,42,43</sup>, Emiliano Dalla<sup>40</sup>, Carrie A. Davis<sup>44</sup>, Michael Detmar<sup>45</sup>, Alexander D. Diehl<sup>46</sup>, Taeko Dohi<sup>47</sup>, Finn Drablos<sup>48</sup>, Albert S. B. Edge<sup>49</sup>, Matthias Edinger<sup>4,5</sup>, Karl Ekwall<sup>50</sup>, Mitsuhiro Endoh<sup>51,52</sup>, Hideki Enomoto<sup>53</sup>, Michela Fagioli<sup>5,4</sup>, Lynsey Fairbairn<sup>5</sup>, Hai Fang<sup>55</sup>, Mary C. Farach-Carson<sup>56</sup>, Geoffrey J. Faulkner<sup>57</sup>, Alexander V. Favorov<sup>51,58,59</sup>, Malcolm E. Fisher<sup>6</sup>, Martin C. Frith<sup>60</sup>, Rie Fujita<sup>61</sup>, Shiro Fukuda<sup>1</sup>, Cesare Furlanello<sup>2,1</sup>, Masaaki Furuno<sup>2</sup>, Jun-ichi Furusawa<sup>51,62,63</sup>, Teunis B. Geijtenbeek<sup>64</sup>, Andrew P. Gibson<sup>64</sup>, Thomas Gingeras<sup>65</sup>, Daniel Goldowitz<sup>65</sup>, Julian Gough<sup>66</sup>, Sven Guhl<sup>67</sup>, Reto Guler<sup>31,32</sup>, Stefano Gustinich<sup>68</sup>, Thomas J. Ha<sup>69</sup>, Masahide Hamaguchi<sup>70</sup>, Mitsuho Hara<sup>68</sup>, Matthias Harbers<sup>71</sup>, Jayson Harshbarger<sup>7,2</sup>, Akira Hasegawa<sup>1,2</sup>, Yuki Hasegawa<sup>1,2</sup>, Takehiro Hashimoto<sup>7</sup>, Meenhard Herlyn<sup>69</sup>, Kelly J. Hitchens<sup>25,26</sup>, Shannan J. Ho Sui<sup>16</sup>, Oliver M. Hofmann<sup>16</sup>, Ilka Hoof<sup>1,2</sup>, Fumi Hori<sup>1,2</sup>, Lukasz Huminiński<sup>17</sup>, Kei Iida<sup>70</sup>, Tomokatsu Ikawa<sup>51,52</sup>, Boris R. Jankovic<sup>1,4</sup>, Hui Jia<sup>72</sup>, Anagha Joshi<sup>6</sup>, Giuseppe Jurman<sup>21</sup>, Bogumil Kaczkowski<sup>1,2</sup>, Chieko Kai<sup>73</sup>, Kaoru Kaida<sup>1,2</sup>, Ai Kaiho<sup>74</sup>, Kazuhiro Kajiyama<sup>1,2</sup>, Mutsuomi Kanamori-Katayama<sup>1</sup>, Artem S. Kasianov<sup>10</sup>, Takeya Kasukawa<sup>2</sup>, Shintaro Katayama<sup>1</sup>, Sachi Kato<sup>1,2</sup>, Shuji Kawaguchi<sup>70</sup>, Hiroshi Kawamoto<sup>51</sup>, Yuki I. Kawamura<sup>47</sup>, Tsugumi Kawashima<sup>1,2</sup>, Judith S. Kempfle<sup>49</sup>, Tony J. Kenna<sup>29</sup>, Juha Kere<sup>50,75</sup>, Levon M. Khachigian<sup>74</sup>, Toshio Kitamura<sup>75</sup>, S. Peter Klinken<sup>76</sup>, Alan J. Knox<sup>77</sup>, Miki Kojima<sup>1,2</sup>, Soichi Kojima<sup>68</sup>, Naoto Kondo<sup>1,2</sup>, Haruhiko Koseki<sup>51,52</sup>, Shigeo Koyasu<sup>51,52,62</sup>, Sarah Krampitz<sup>45</sup>, Atsuta Kubosaki<sup>1</sup>, Andrew T. Kwon<sup>1,2</sup>, Jerome F. J. Laros<sup>54</sup>, Weonju Lee<sup>78</sup>, Andreas Lennartsson<sup>50</sup>, Kang Li<sup>31</sup>, Berit Lilje<sup>31</sup>, Leonard Lipovich<sup>71</sup>, Alan Mackay-sim<sup>79</sup>, Ri-ichiro Manabe<sup>1,2</sup>, Jessica C. Mar<sup>39</sup>, Benoit Marchand<sup>44</sup>, Anthony Mathelier<sup>65</sup>, Niklas Meijer<sup>22</sup>, Alison Meyneir<sup>18</sup>, Yosuke Mizuno<sup>60</sup>, David A. de Lima Morais<sup>61</sup>, Hiromasa Morikawa<sup>67</sup>, Mitsuuru Morimoto<sup>53</sup>, Kazuhiro Moro<sup>51,52,62,82</sup>, Ethymios Motakis<sup>1,2</sup>, Hozumi Motohashi<sup>83</sup>, Christine L. Mummeny<sup>84</sup>, Mitsuhiro Murata<sup>1,2</sup>, Sayaka Nagao-Sato<sup>1</sup>, Yutaka Nakach<sup>85,86</sup>, Fumio Nakahara<sup>70</sup>, Toshiyuki Nakamura<sup>72</sup>, Yukio Nakamura<sup>70</sup>, Kenichi Nakazato<sup>1</sup>, Erik van Nimwegen<sup>87</sup>, Noriko Ninomiya<sup>88</sup>, Hiromi Nishiyori<sup>1,2</sup>, Shohei Noma<sup>1,2</sup>, Tadasu Nozaki<sup>89</sup>, Soichi Ogishima<sup>90</sup>, Naganori Ohkura<sup>91</sup>, Hiroko Ohmura<sup>1,2,4</sup>, Hiroshi Ohno<sup>1,52</sup>, Mitsuhiro Ohshima<sup>92</sup>, Mariko Okada-Hatakeyama<sup>31,50</sup>, Yasushi Okazaki<sup>90,95</sup>, Valerio Orlando<sup>30,37</sup>, Dmitry A. Ovchinnikov<sup>25</sup>, Arnab Pain<sup>4,37</sup>, Robert Passier<sup>94</sup>, Margaret Patrikakis<sup>95</sup>, Helena Persson<sup>50</sup>, Silvano Piazza<sup>40</sup>, James G. D. Prendergast<sup>1,2</sup>, Owen J. L. Rackham<sup>96</sup>, Jordan A. Ramilowski<sup>1,2</sup>, Mamoon Rashid<sup>14,37</sup>, Timothy Ravasi<sup>37,38</sup>, Patrizia Rizzu<sup>19</sup>, Marco Roncador<sup>21</sup>, Sugata Roy<sup>1,2</sup>, Morten B. Rye<sup>48</sup>, Eri Saijyo<sup>1</sup>, Antti Sajantila<sup>90</sup>, Akiho Saka<sup>1</sup>, Shimon Sakaguchi<sup>67</sup>, Mizuho Sakai<sup>1,2</sup>, Hiroki Sato<sup>72</sup>, Hironori Satoh<sup>91</sup>, Suzana Savvi<sup>3,32</sup>, Alka Saxena<sup>1,†</sup>, Claudio Schneider<sup>40,91</sup>, Erik A. Schultes<sup>64</sup>, Gundula G. Schulze-Tanzil<sup>92</sup>, Anita Schwegmann<sup>31,32</sup>, Thierry Sengstag<sup>3</sup>, Guojun Sheng<sup>33</sup>, Hisashi Shimoji<sup>1</sup>, Yishai Shimon<sup>35</sup>, Jay W. Shin<sup>1,2</sup>, Christophe Simon<sup>1,2</sup>, Daisuke Sugiyama<sup>93</sup>, Takaaki Sugiyama<sup>92</sup>, Masanori Suzuki<sup>1</sup>, Naoko Suzuki<sup>1,2</sup>, Rolf K. Swoboda<sup>99</sup>, Peter A. C. 't Hoen<sup>64</sup>, Michihira Tagami<sup>1,2</sup>, Naoko Takahashi<sup>1,2</sup>, Jun Taka<sup>61</sup>, Hiroshi Tanaka<sup>88</sup>, Hideki Tatsukawa<sup>94</sup>, Zuoqian Tatam<sup>64</sup>, Mark Thompson<sup>64</sup>, Hiroo Toyoda<sup>87</sup>, Tetsuro Toyoda<sup>70</sup>, Eivind Valen<sup>95</sup>, Marc van de Wetering<sup>41</sup>, Linda M. van den Berg<sup>63</sup>, Roberto Verardo<sup>40</sup>, Dipri Vijayan<sup>25,26</sup>, Ilya E. Vorontsov<sup>1</sup>, Wyeth W. Wasserman<sup>1</sup>, Shoko Watanabe<sup>1</sup>, Christine A. Wells<sup>25,26</sup>, Louise N. Wintneringham<sup>75</sup>, Ernst Wolvetang<sup>25</sup>, Emily J. Wood<sup>1</sup>, Yoko Yamaguchi<sup>96</sup>, Masayuki Yamamoto<sup>61</sup>, Misako Yoneda<sup>1</sup>, Yohei Yokura<sup>93</sup>, Shigehiro Yoshida<sup>1</sup>, Susan E. Zabierowski<sup>99</sup>, Peter G. Zhang<sup>63</sup>, Xiaobei Zhao<sup>1</sup>, Silvia Zucchelli<sup>96</sup>, Kim M. Summers<sup>9</sup>, Harukazu Suzuki<sup>1,2</sup>, Carsten O. Daub<sup>1</sup>, Jun Kawai<sup>1,3</sup>, Peter Heutink<sup>1</sup>, Winston Hide<sup>16</sup>, Tom C. Freeman<sup>9</sup>, Boris Lenhard<sup>80,97</sup>, Vladimir B. Bajic<sup>1</sup>, Martin S. Taylor<sup>98</sup>, Vsevolod J. Makeev<sup>3,103,98</sup>, Albin Sandelin<sup>1</sup>, David A. Hume<sup>9</sup>, Piero Carninci<sup>1,2</sup>, Yoshihide Hayashizaki<sup>1,3</sup>
- <sup>1</sup>RIKEN Omics Science Center (OSC), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan. <sup>2</sup>RIKEN Center for Life Science Technologies (Division of Genomic Technologies) (CLST (DGT)), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. <sup>3</sup>RIKEN Preventive Medicine and Diagnosis Innovation Program (PMI), 2-1 Hirasawa, Wako-shi, Saitama 351-0198, Japan. <sup>4</sup>Department of Internal Medicine III, University Hospital Regensburg, F.-J.-Strauss Allee 11, D-93042 Regensburg, Germany. <sup>5</sup>Regensburg Centre for Interventional Immunology (RCI), D-93042 Regensburg, Germany. <sup>6</sup>The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Edinburgh, Midlothian EH25 9RG, UK. <sup>7</sup>Department of Biology, University of Bergen, Thormøhlensgate 53, NO-5006 Bergen, Norway. <sup>8</sup>Faculty of Medicine, Institute of Clinical Sciences, MRC Clinical Sciences Centre, Imperial College London, Hammersmith Hospital Campus, London W12 0NN, UK. <sup>9</sup>Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Vavilovstr. 32, Moscow 119991, Russia. <sup>10</sup>Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkin str. 3, Moscow 119991, Russia. <sup>11</sup>The Bioinformatics Centre, Department of Biology and BRIC, University of Copenhagen, Ole Maaloes Vej 5, DK 2200 Copenhagen, Denmark. <sup>12</sup>Genomics Division, Lawrence Berkeley National Laboratory, 84R01, 1 Cyclotron Road, Berkeley, California 94720, USA. <sup>13</sup>Mouse Informatics, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. <sup>14</sup>Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Ibn Al-Haytham Building -2, Thuwal 23955-6900, Kingdom of Saudi Arabia. <sup>15</sup>Institute of Natural and Mathematical Sciences, Massey University, Private Bag 102-904, North Shore Mail Centre, 0745 Auckland, New Zealand. <sup>16</sup>Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave, Boston, Massachusetts 02115, USA. <sup>17</sup>Department of Cell and Molecular Biology, Karolinska Institutet, P.O. Box 285, SE-171 77 Stockholm, Sweden. <sup>18</sup>MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine (MRC-IGMM), University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK. <sup>19</sup>Department of Clinical Genetics, VU University Medical Center Amsterdam, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands. <sup>20</sup>Graduate Program in Areas of Basic and Applied Biology, Abel Salazar Biomedical Sciences Institute, University of Porto, Rua de Jorge Viterbo Ferreira n. 228, 4050-313 Porto, Portugal. <sup>21</sup>Predictive Models for Biomedicine and Environment, Fondazione Bruno Kessler, via Sommarive 18, 38123 Trento, Italy. <sup>22</sup>Department of Medicine, Karolinska Institutet at Karolinska University Hospital, Huddinge, SE-141 86 Huddinge, Sweden. <sup>23</sup>Department of Dermatology and Allergy, Charité Campus Mitte, Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany. <sup>24</sup>Biozentrum, University of Basel, Klingelbergstrasse 50-70, 4056 Basel, Switzerland. <sup>25</sup>Australian Institute for Bioengineering and Nanotechnology (AIBN), University of Queensland, Brisbane St Lucia, Queensland 4072, Australia. <sup>26</sup>Australian Infectious Diseases Research Centre (AID), University of Queensland, Brisbane St Lucia, Queensland 4072, Australia. <sup>27</sup>Department of Biological Sciences, University of Delaware, Newark, Delaware 19713, USA. <sup>28</sup>Bioinformatics and Computational Biology, The Jackson Laboratory, 600 Main Street, Bar Harbor, Maine 04609, USA. <sup>29</sup>Diamantina Institute, University of Queensland, Brisbane St Lucia, Queensland 4072, Australia. <sup>30</sup>IRCCS Fondazione Santa Lucia, via del Fosso di Fiorano 64, 00143 Rome, Italy. <sup>31</sup>Immunology and Infectious Disease, International Centre for Genetic Engineering & Biotechnology (ICGEB) Cape Town component, Anzio Road, Observatory 7925, Cape Town, South Africa. <sup>32</sup>Division of Immunology, Institute of Infectious Diseases and Molecular Medicine (IDM), University of Cape Town, Anzio Road, Observatory 7925, Cape Town, South Africa. <sup>33</sup>Department of Systems Biology, Columbia University Medical Center, 1130 St. Nicholas Avenue, New York, New York 10032, USA. <sup>34</sup>Department of Biochemistry and Molecular Biophysics, Columbia University Medical Center, 701 West 168th Street, New York, New York 10032, USA. <sup>35</sup>Department of Biomedical Informatics, Columbia University Medical Center, 622 West 168th Street, VC5, New York, New York 10032, USA. <sup>36</sup>Institute of Cancer Genetics, Columbia University Medical Center, Herbert Irving Comprehensive Cancer Center, 1130 St. Nicholas Avenue, New York, New York 10032, USA. <sup>37</sup>Biological and Environmental Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Ibn Al-Haytham Building -2, Thuwal 23955-6900, Kingdom of Saudi Arabia. <sup>38</sup>Applied Mathematics and Computational Science Program, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia. <sup>39</sup>Department of Systems and Computational Biology, Albert Einstein College of Medicine, The Bronx, New York, New York 10461, USA. <sup>40</sup>Laboratorio Nazionale del Consorzio Interuniversitario per le Biotecnologie (LNCIB), Padriciano 99, 34149 Trieste, Italy. <sup>41</sup>Hubrecht Institute, Uppsalaalan 8, 3584 CT Utrecht, The Netherlands. <sup>42</sup>The Royal Netherlands Academy of Arts and Sciences, P.O. Box 19121, NL-1000 GC Amsterdam, The Netherlands. <sup>43</sup>University Medical Center Utrecht, Postbus 85500, 3508 GA Utrecht, The Netherlands. <sup>44</sup>Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11797, USA. <sup>45</sup>Institute of Pharmaceutical Sciences, ETH Zurich, Vladimir-Pregel-Weg 3, HCI H 303, 8093 Zurich, Switzerland. <sup>46</sup>Department of Neurology, University at Buffalo School of Medicine and Biomedical Sciences, New York State Center of Excellence in Bioinformatics and Life Sciences, 701 Ellicott Street, Buffalo, New York 14203, USA. <sup>47</sup>Gastroenterology, Research Center for Hepatitis and Immunology Research Institute, National Center for Global Health and Medicine, 1-7-1 Kohnodai, Ichikawa, Chiba 272-8516, Japan. <sup>48</sup>Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), P.O. Box 8905, NO-7491 Trondheim, Norway. <sup>49</sup>Department of Otolaryngology and Laryngology, Harvard Medical School, Massachusetts Eye and Ear Infirmary, Eaton-Peabody Lab, 243 Charles Street, Boston, Massachusetts 02114, USA. <sup>50</sup>Department of Biosciences and Nutrition, Center for Biosciences, Karolinska Institutet, Hälsovägen 7-9, SE-141 83 Huddinge, Sweden. <sup>51</sup>RIKEN Research Center for Allergy and Immunology (RCAI), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. <sup>52</sup>RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. <sup>53</sup>RIKEN Center for Developmental Biology (CDB), 2-2-3 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan. <sup>54</sup>FM Kirby Neurobiology Center, Children's Hospital Boston, Harvard Medical School, 300 Longwood Avenue, Boston, Massachusetts 02115, USA. <sup>55</sup>Department of Computer Science, University of Bristol, Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, UK. <sup>56</sup>Department of Biochemistry and Cell Biology, Rice University, Houston, Texas 77251-1892, USA. <sup>57</sup>Cancer Biology Program, Mater Medical Research Institute, Raymond Terrace, South Brisbane, Queensland 4101, Australia. <sup>58</sup>Department of Oncology, Division of Oncology, Biostatistics and Bioinformatics, Johns Hopkins University School of Medicine, 550 North Broadway, Baltimore, Maryland 21205, USA. <sup>59</sup>State Research Institute of Genetics and Selection of Industrial Microorganisms GosNIIgenetika, 1-st Dorozhny pr., 1, 117545 Moscow, Russia. <sup>60</sup>Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan. <sup>61</sup>Department of Medical Biochemistry, Tohoku University Graduate School of Medicine, 2-1 Seiryō-machi, Aoba-ku, Sendai, Miyagi 980-8575, Japan. <sup>62</sup>Department of Microbiology and Immunology, Keio University School of Medicine, 35 Shinanomachi, Shinjuku, Tokyo 160-8582, Japan. <sup>63</sup>Experimental Immunology, Academic Medical Center, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands. <sup>64</sup>Department of Human Genetics, Leiden University Medical Center, Einthovenweg 20, 2333 ZC Leiden, The Netherlands. <sup>65</sup>Department of Medical Genetics, Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, University of British Columbia, 950 West 28th Avenue, Vancouver, British Columbia V5Z 4H4, Canada. <sup>66</sup>Neuroscience, SISSA, via Bonomea 265, 34136 Trieste, Italy. <sup>67</sup>Experimental Immunology, Immunology Frontier Research Center, Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan. <sup>68</sup>RIKEN Advanced Science Institute (ASI), 2-1 Hirasawa, Wako, Saitama 351-0198, Japan. <sup>69</sup>Melanoma Research Center, The Wistar Institute, 3601 Spruce Street, Philadelphia, Pennsylvania 19104, USA. <sup>70</sup>RIKEN Bioinformatics And Systems Engineering Division (BASE), 1-7-22 Suehiro, Tsurumi, Yokohama, Kanagawa 230-0045, Japan. <sup>71</sup>Center for Molecular Medicine and Genetics, Wayne State University, 3228 Scott Hall, 540 East Canfield Street, Detroit, Michigan 48201-1928, USA. <sup>72</sup>Laboratory Animal Research Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan. <sup>73</sup>Science for Life Laboratory, Box 1031, SE-171 21 Solna, Sweden. <sup>74</sup>Centre for Vascular Research, University of New South



Wales, Sydney, New South Wales 2052, Australia.<sup>75</sup> Division of Cellular Therapy and Division of Stem Cell Signaling, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan.<sup>76</sup> Harry Perkins Institute of Medical Research, and the Centre for Medical Research, University of Western Australia, QO Block, QEII Medical Centre, Nedlands, Perth, Western Australia 6009, Australia.<sup>77</sup> Respiratory Medicine, University of Nottingham, Clinical Sciences Building, City Hospital, Hucknall Road, Nottingham NG5 1PB, UK.<sup>78</sup> Department of Dermatology, Kyungpook National University School of Medicine, 130 Dongdeok-ro Jung-gu, Daegu 700-721, South Korea.<sup>79</sup> National Centre for Adult Stem Cell Research, Eskitis Institute for Cell and Molecular Therapies, Griffith University, Brisbane, Queensland 4111, Australia.<sup>80</sup> Division of Functional Genomics and Systems Medicine, Research Center for Genomic Medicine, Saitama Medical University, 1397-1 Yamane, Hidaka, Saitama 350-1241, Japan.<sup>81</sup> Faculty of Engineering, University of Bristol, Merchant Venturers Building, Woodland Road, Clifton BS8 1UB, UK.<sup>82</sup> PRESTO, Japanese Science and Technology Agency (JST), 7 Gobancho, Chiyodaku, Tokyo 102-0076, Japan.<sup>83</sup> Center for Radioisotope Sciences, Tohoku University Graduate School of Medicine, 2-1 Seiryō-machi, Aoba-ku, Sendai, Miyagi 980-8575, Japan.<sup>84</sup> Anatomy and Embryology, Leiden University Medical Center, Einthovenweg 20, P.O. Box 9600, 2300 RC Leiden, The Netherlands.<sup>85</sup> Division of Translational Research, Research Center for Genomic Medicine, Saitama Medical University, 1397-1 Yamane, Hidaka, Saitama 350-1241, Japan.<sup>86</sup> RIKEN BioResource Center (BRC), Koyadai 3-1-1, Tsukuba, Ibaraki 305-0074, Japan.<sup>87</sup> Department of Clinical Molecular Genetics, School of Pharmacy, Tokyo University of Pharmacy and Life Sciences, 1432-1 Horinouchi, Hachioji, Tokyo 192-0392, Japan.<sup>88</sup> Department of Bioinformatics, Medical Research Institute, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan.<sup>89</sup> Department of Biochemistry, Ohu University School of Pharmaceutical Sciences, Misumido 31-1, Tomitamachi, Koriyama, Fukushima 963-8611, Japan.<sup>90</sup> Hjelte Institute, Department of Forensic Medicine, University of Helsinki, Kytösuntie 11, 00300 Helsinki, Finland.<sup>91</sup> DSMB Dipartimento Scienze Mediche e Biologiche University of Udine, P.le Kolbe 3, 33100 Udine, Italy.<sup>92</sup> Department of Orthopedic, Trauma and Reconstructive Surgery, Charité Universitätsmedizin Berlin, Garystrasse 5, 14195 Berlin, Germany.<sup>93</sup> Center for Clinical and Translational Research, Kyushu University Hospital, Station for Collaborative Research1 4F, 3-1-1 Maidashi, Higashi-Ku, Fukuoka 812-8582, Japan.<sup>94</sup> Graduate School of Pharmaceutical Sciences, Nagoya University, Furo-cho, Chikusa, Nagoya, Aichi 464-8601, Japan.<sup>95</sup> Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA.<sup>96</sup> Department of Biochemistry, Nihon University School of Dentistry, 1-8-13, Kanda-Surugadai, Chiyoda-ku, Tokyo 101-8310, Japan.<sup>97</sup> Department of Informatics, University of Bergen, Hagtekologisenteret, Thormøhlensgate 53, NO-5008 Bergen, Norway.<sup>98</sup> Department of Biological and Medical Physics, Moscow Institute of Physics and Technology (MIPT) 9, Institutsky Per., Dolgoprudny, Moscow Region 141700, Russia.  
 †Present addresses: Institute of Predictive and Personalized Medicine of Cancer, Ctra. de Can Roti, camí de les escoles, s/n, 08916 Badalona (Barcelona), Spain (Y.A.M.); Biomedical Cybernetics Group, Biotechnology Center (BIOTEC), Technische Universität Dresden, Dresden, Germany (C.V.C.); Genomics Core Facility, Biomedical Research Centre, Guy's Hospital, London SE1 9RT, UK (A. Saxena); RIKEN Advanced Center for Computing and Communication (ACCC), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045 Japan (H. Ohmura); Research Center for Molecular Medicine of the Austrian Academy of Sciences (CeMM), 1090 Vienna, Austria (C. Schmidt); Department of Biological and Biomedical Sciences, Harvard University, Cambridge, Massachusetts 02138, USA (J.B.); Department of Bioclinical Informatics, Tohoku Medical Megabank Organization, Tohoku University, Sendai 980-8573, Japan (S.O.).

\*These authors contributed equally to this work.



# Chapter 4

## A high resolution spatial promoterome of the human brain

Francescato M\*, Vitezic M\*, Rizzu P, Simón-Sánchez J, Andersson R, Kawaji H, Itoh M, Kondo N, Lassmann T, Kawai J, Suzuki H, Hayashizaki Y, Daub CO, Sandelin A, de Hoon MJL, Carninci P, Forrest ARR, Heutink P and the FANTOM consortium. 2014. A high resolution spatial promoterome of the human brain. Manuscript.



## Abstract

The human CNS is an extremely complex organ that governs our abilities for cognition, reasoning and emotions and is the control center for the body. Its morphology and functionality during development have been well studied, but the molecular mechanisms contributing to its function and maintenance later in life remain poorly understood. Complexity at the transcriptional level is likely to play a major role in defining its morphological and functional characteristics. To investigate this we used single molecule Cap Analysis of Gene Expression to create a high-resolution atlas of transcription start sites for 15 anatomical regions of the human central nervous system, using post mortem samples derived from three aged adult donors. Sequencing on average 5 million reads per sample, we identified 95912 CAGE-defined tag clusters (TCs), supporting the expression of 19018 genes. Using the largest tissue collection produced to date with a uniform platform we show that the CNS has a unique expression signature, not limited to protein coding genes but extending to lncRNAs and novel transcripts. Additionally, it is distinguished by a significantly higher transcriptional complexity. We show that transcripts up-regulated in brain arise in a specific transcriptional context, being more often transcribed from CG rich regions, simple and low complexity repeats. We identify a set of 183 transcription factors and 206 lncRNAs up-regulated in brain which co-expression patterns identify super-groups of regions with related function/developmental derivation. 9758 TCs are differentially expressed across regions, representing four major co-expression groups, each of which includes genes that are known to be relevant for the function of the associated regions. E.g. TBR1 and ARNT2 transcription factors and cortex markers FXYD6, CCK and CBLN2 belong to the co-expression group associated with cortex and limbic system. In addition we find in this group 37 lncRNAs of unknown function and 147 intergenic TCs, over 74% of which overlap frontal-derived H3K4me3 ChIP-seq data, strongly supporting that they correspond to genuine novel transcripts. Due to its high-resolution and the large variety of CNS regions represented, this study provides an invaluable resource for understanding region-specific transcriptional regulation and provide testable hypotheses that can be followed up in the

laboratory.

## **Introduction**

The human brain is an exceptionally sophisticated organ divided into distinct anatomical districts that are characterized by specific cellular compositions and functions and are interconnected by intricate communication networks. Complexity at the transcriptional level is likely to play a major role in the establishment and maintenance of the morphological and functional complexity of the brain and its multifaceted parts. Studies to date, investigating genome wide expression profiles of the human central nervous system across different regions and developmental stages (Roth et al. 2006; Kang et al. 2011; Colantuoni et al. 2011; Hawrylycz et al. 2012) have provided invaluable insight into the transcriptional dynamics and regulation in different areas of the human brain. However, these studies mostly relied on array-based technologies that are biased in their probe design and limited by their inability to detect novel transcripts and transcript isoforms, or to distinguish between closely related paralogous sequences. In addition they often cannot quantify absolute expression (Fu et al. 2009). Next generation sequencing (NGS) is rapidly replacing microarrays as the technique of choice for transcription profiling studies in an effort to overcome these limitations. Cap analysis of gene expression (CAGE) is a transcriptome exploration technology that captures the 5' end of capped RNA transcripts (Kodzius et al. 2006; Takahashi et al. 2012) allowing for the high resolution profiling of transcription start sites (TSSs) in a quantitative and annotation-independent manner. CAGE has been successfully employed to profile transcription in several organisms and clonal cell lines in varying experimental conditions (FANTOM consortium 2009; Hoskins et al. 2011; Plessy et al. 2012; ENCODE Project Consortium 2012) giving novel insights into mammalian transcriptional regulation (Lenhard et al. 2012) and has been the technology of choice of the FANTOM (Functional Annotation of the Mammalian Genome) consortium (FANTOM Consortium 2009). The FANTOM5 project uses CAGE, adapted to the single-molecule sequencer Heliscope (Kanamori-Katayama

et al. 2011) to avoid additional PCR steps and improve its quantitiveness, to profile over 900 human tissues, primary cells and cell lines, aiming to build a complete promoter map to uncover the transcriptional regulatory networks defining every human primary cell type (Forrest et al. 2014). As part of FANTOM5, we profiled transcription for 15 regions of the human central nervous system (CNS, Table 1, Supplementary Figure 1), using post mortem tissue from three aged adult donors. The regions belong to distinct anatomical and functional domains and are involved in a wide range of neurological phenotypes, including major diseases. Our data extend and complement microarray-based brain gene expression studies (Roth et al. 2006; Kang et al. 2011; Colantuoni et al. 2011; Hawrylycz et al. 2012), the recently published ENCODE data (primarily based upon a limited set of clonal cell lines (ENCODE Consortium 2012)) and our previous work (based on CAGE profiling of a set of five brain regions (Pardo et al. 2013)).

This study provides an important resource for in depth brain specific functional annotation. Using the largest collection produced to date with a uniform platform we show that brain has a distinctive expression signature with respect to other tissues, not limited to protein coding genes but extending to lncRNAs and novel transcripts. Additionally, it is distinguished by a higher transcriptional complexity, to which non-coding transcripts importantly contribute. We show that transcripts up-regulated in brain are characterized by a specific transcriptional context, being often derived from CG rich regions and specific classes of repeats. We also identify a set of transcription factors and lncRNAs up-regulated in brain, that might have an important role in brain-specific transcriptional regulation. We assess the extent of regionally biased transcription across distinct regions of the adult brain, highlighting a set of locally expressed lncRNAs and transcription factors. This work is part of the FANTOM5 project. Data downloads, genomic tools and co-published manuscripts are summarized at <http://fantom.gsc.riken.jp/5/>.

## Results

### ***The complex transcriptome of the human brain***

We identified 95,912 tag clusters (TCs) expressed in the human CNS, 95.0% of which could be associated to 19,018 GENCODE genes. 78.9% and 16.9% of the annotated TCs mapped to protein-coding and non-coding transcripts respectively. The most represented non-coding biotypes were processed transcripts (47.0%), retained introns (23.9%) and long non-coding RNAs (lncRNAs) (8.7%) (Table 2). We also identified 4,779 (5.0%) TCs mapping to previously un-annotated, intergenic regions, representing bona fide novel transcripts. Using the publicly available ChIP-seq dataset published in (Shulha et al. 2013), which reports genome wide maps for the histone H3K4me3 (associated with promoters that are active or poised to be activated (Barski et al. 2007)) in nuclei collected from prefrontal cortex, we found that 34.5% of these intergenic TCs overlap H3K4me3 signature, supporting the hypothesis that they represent TSSs of novel transcripts. Using the FANTOM5 tissue collection and the advantage of having a broad set of samples profiled with the same technology (35 CNS tissues and 91 heterogeneous non-CNS tissues, listed in Supplementary Table 1), we were able to investigate overall expression differences between CNS and other tissues. We produced multidimensional scaling (MDS) plots of CAGE expression profiles for four subsets of the data, representing coding genes, transcription factors, lncRNAs and intergenic TCs. As shown in Figure 1 (panels a. to d.) for all the four groups the CNS tissues clustered together and were clearly separated from the other tissues, showing that the CNS expresses a specific range of coding genes (in particular transcription factors) and lncRNAs, and is also distinguished by the expression patterns of putative novel transcripts. Additionally, analogous clustering based on transcribed enhancers (Andersson et al. 2014, Figure 3) similarly showed that CNS samples clearly separate from other FANTOM5 human tissues. These results show that the CNS expression signature is distinctive with respect to other tissues, not only on the level of coding genes but also on lncRNAs, novel transcripts and transcriptional regulators, such as transcription factors and enhancers. This is particularly interesting in light of the fact that although the non-



CNS tissues are extremely heterogeneous, they form an homogeneous cluster with respect to CNS samples.

One hypothesis to explain this remarkable separation is that brain tissues express a broader range of transcripts, therefore inducing the expression patterns observed. To assess this we examined the cumulative distribution of tags accounted for by the 10,000 most highly expressed TCs in each tissue library: as shown in Figure 1e in general the curves that represent CNS samples grow slower than the ones representing other tissues, suggesting that brain has a more complex and diversified transcriptome. To quantify this, we calculated the number of TCs required to cover at least 50% of the tags sequenced in each of the libraries, similarly to what was described in (Jongeneel et al. 2005). These numbers, referred to as N50, can be considered as a measure of transcriptional complexity, since tissues with simple transcriptional programs are characterized by a low N50 value (e.g. in the library prepared from salivary gland the two most highly expressed TCs, mapping to the genes Submaxillary Gland Androgen Regulated Protein 3B and Statherin, accounted for almost 60% of the total tags sequenced in that library, therefore N50 for salivary gland was 1). Comparing N50 values for the CNS samples against the other tissues, we observed a significant difference ( $p$ -value =  $1.213e-09$ , Wilcoxon rank sum test, Figure 1f) showing that in general CNS samples give rise to more complex libraries. Repeating this analysis based only on TCs mapping to non-coding loci we still obtained a significant difference between CNS and other tissues (Supplementary Figure 2), showing that the non-coding fraction contributes significantly to the complexity of brain transcriptome.

### ***Transcripts up-regulated in brain***

To identify TCs with higher expression in the CNS with respect to other tissues, we performed differential expression analysis. Of the 152,952 TCs expressed in the FANTOM5 tissue collection, 55,033 (36.0%) were differentially expressed; in particular, one third had higher expression in brain (18,626 TCs, mapping to 3,928 distinct genes; genomic coordinates, fold-changes and  $p$ -values provided in Supplementary Table 2) (Figure 2a) and will be referred to as Brain-up in the rest

of the manuscript, to indicate that they are up-regulated in brain.

We first investigated the general differences in genomic context between the two sets of TCs, up-regulated and down-regulated in brain tissues. Comparing GENCODE annotations gave similar proportions of known TSS, coding sequences, antisense or intergenic signals. However TCs with higher expression in brain were slightly more likely to be distal to annotated TSSs (between 500 and 1000 bps upstream) or 3'UTR-derived (Supplementary Figure 4). As deep sequencing of the human brain transcriptome has not been done previously to the same extent, these differences may reflect transcripts up-regulated in CNS that currently lack accurate gene models. It was previously reported that genes expressed in brain are frequently located in CpG rich regions (Roider et al. 2009): consistently with this observation, 5,811 (31.2%) of the brainUp TCs were located in CpG islands, as opposed to 7,642 (21.0%) in the down-regulated group ( $p < e-16$ , Fisher exact test). It has also been suggested that brain tissues express a distinctive repertoire of repeats (Faulkner et al. 2009; Xu et al. 2010; Tyekucheva et al. 2011): in general we observed a slightly larger proportion of brainUp TCs overlapping repeats (18.9% vs. 13.1%). In particular, a remarkably large proportion of brainUp TCs overlapped simple and low complexity repeats (59.0% of brainUp TCs, as opposed to 26.5%), while there were less expressed LTRs (8.8% as opposed to 34.4%, Figure 2b). There was no overall difference in the number of expressed Long and Short Interspersed Elements (LINE and SINE respectively); however a significant difference was observed in the relative proportion of expressed Alu (in the SINE family) and L1 (in the LINE family) repeats (Figure 2c). Genes containing TCs up-regulated in brain were highly enriched in GO Biological Process terms and KEGG pathways related to brain function (Figure 3 a and b) as well as Genetic Association Database diseases such as schizophrenia, epilepsy and alcohol dependence (full list of enrichments provided in Supplementary Table 3). Looking at candidates that might have an important impact at the regulatory level, we investigated brainUp transcription factors and lncRNAs, with the specific aim of identifying novel elements that could be of interest for future research on genes involved in the establishment and/or maintenance of CNS transcriptional specificity. We identified 520 brainUp TCs mapping to transcription factors (183

genes): these included several examples of genes with critical roles in the development of the CNS such as *TBR1*, required for early cortical development (Bulfone et al. 1995), *ZIC1* and *ZIC4*, fundamental for cerebellar development (Blank et al. 2011), *BHLHE22*, involved in neocortex development (Joshi et al. 2008). The most highly expressed TF was *TSC22D4* (Figure 3c), suggested to be important for granule cells differentiation in mouse (Canterini et al. 2012). Importantly, we also found several brainUp TFs with unknown function, such as the poorly characterized Zinc Finger proteins *ZNF25*, *ZNF273*, *ZNF302* and several others (the full list of brainUp TCs mapping to TFs ranked by expression in the CNS is provided in Supplementary Table 4). Interestingly some of these have recently been shown to be relevant for major diseases: *HIVEP3*, shown to be an essential regulator of adult bone formation (Jones et al. 2006) was suggested as candidate gene for the *PARK10* locus associated to Parkinson's Disease (Li et al. 2007); *PRDM8* was recently indicated as the causal protein of the early onset Lafora disease, a type of progressive myoclonus epilepsy (Turnbull et al. 2012), *ZNF385D* was recently linked to reading disability and language impairment (Eicher et al. 2013) and negative symptoms in schizophrenia (Xu et al. 2013); *TEF*, associated with sleep disturbances and depression in Parkinson's disease patients (Hua et al. 2012; Hua et al. 2012). We identified 419 brainUp TCs mapping to 206 distinct lncRNAs (full list of brainUp TCs mapping to lncRNAs ranked by expression in the CNS is provided in Supplementary Table 4). Only two of them corresponded to known genes: the maternally imprinted genes *MEG3* and *H19*. The remaining lncRNAs, such as the most highly expressed in brain (*AC073479.1*, Figure 3d) had no known annotation; intriguingly, however, unsupervised clustering based on their expression profiles identified four groups of regions with related developmental derivation, function and/or projections (Figure 3e): 1) cerebellum, 2) cortex along with amygdala and hippocampus (cortex-limbic system group), 3) caudate and putamen (striatum), thalamus, globus pallidus, substantia nigra, locus coeruleus, spinal cord and medulla oblongata (brain stem-basal ganglia group). Similarly, clustering based on expression of brainUp TCs mapping to TFs identified the same four groups of regions (Figure 3f), suggesting that these lncRNAs and TFs up-regulated in brain have functional relevance in the biology of

these regions.

### ***Region specific transcription in the adult***

In order to assess individual differences in expression across distinct brain regions, we performed differential expression analysis and identified 9,758 differentially expressed TCs, mapping to 3,891 genes. The region with the largest number of differentially expressed TCs was cerebellum (Figure 4a), possibly due to the fact that it is characterized by the highest neuron to glia ratio in the CNS (Azevedo et al. 2009). Besides this case and consistently with previous reports (Hawrylycz et al. 2012) we didn't identify expression signatures that univocally define single regions, but rather observed four major expression patterns shared across multiple regions, in a way that mimics what we saw for transcription factors and lncRNAs up-regulated in brain (Figure 4b). Based on this evidence, we used k-means clustering to separate the differentially expressed TCs into four mutually exclusive co-expression modules that we named according to the regions they represent (Table 3, Supplementary Figure 5). As expected, in each of them we found TCs mapping to genes that are known markers for the anatomical groups of regions they represent. For example in the cortex-limbic system group we found the genes *FXYD6*, *CCK* and *CBLN2*, markers for cortex layers 2/3/6 (Zeng et al. 2012); the markers for granule cell progenitors *MEIS1*, *PAX6*, *ZIC1* and *ZIC2* (Salero and Hatten 2007) were consistently assigned to cerebellum; the striatum markers *SST*, *DRD1* and *DRD2* were found in the striatum group. The brain stem - basal ganglia group is clearly the most heterogeneous, however we found in this group enzymes involved in the production of specific products that are only synthesized in some of the regions in this group such as *TH*, *DDC* and *DBH*. Additionally we found in this group important components of myelin such as *PLP*, *MOG* and *MBP*, suggesting enrichment in these regions for glial cell types. Interestingly and consistently with our previous observations on genes up-regulated in brain, specific sets of TFs were expressed in each of the four groups (full information provided in Supplementary Table 5). For example, several members of the *HOX* genes, a highly conserved gene family involved in the definition of antero-posterior patterning during embryonic development, were

assigned to the brain stem – basal ganglia group; *ARNT2*, a member of bHLH-PAS TF family linked to nervous system development and previously described as a key factor in mouse hippocampus gene regulation (Valen et al. 2009) belonged to the cortex-limbic system group. Using STRING (<http://string-db.org/>), a database of protein-protein interactions (PPI) based on genomic context, high-throughput experiments, co-expression and literature (Szklarczyk et al. 2011) we could confirm known interactions for a large number of the TFs in each set (Supplementary Figure 6), which suggests that novel connections in each group will be possibly discovered with future research.

Since regionally biased expression of coding genes has been described extensively (Kang et al. 2011; Hawrylycz et al. 2012), we focused on the expression patterns of TCs mapping to poorly characterized transcript classes and genomic regions. In the set of differentially expressed TCs, 1,769 (18.1%) mapped to non-protein coding transcripts and included different biotypes; the most represented classes, accounting for 85.1% of the non-coding fraction, were processed transcripts, retained introns and lncRNAs (Figure 4c). Examples of lncRNAs with regionally biased expression patterns included the uncharacterized transcripts RP11-307B23.1 (brain stem-basal ganglia), RP11-59J5.1 (cerebellum), MIR7-3HG, AC113617.1 and RP11-60A8 (cortex-limbic system), CTA-929C8 (striatum) (the full list of region specific TCs mapping to lncRNAs and corresponding annotations is provided in Supplementary Table 6). Of the differentially expressed TCs mapping to poorly characterized transcript classes, 544 (5.6%) were intergenic, indicating potential new coding/non-coding genes and/or alternative TSSs. Using the histone H3K4me3 ChIP-seq dataset published in (Shulha et al. 2013), we found that 49.1% of these intergenic TCs overlaps the H3K4me3 signature, percentage that increased to 74.1% when restricting to intergenic TCs belonging to the cortex-limbic system group, supporting the hypothesis that they mark TSSs of novel transcripts. A very interesting example of how these intergenic signals can represent novel transcripts is shown in Figure 5d. We identified a set of cerebellum-specific TCs in an intergenic region located 850 kb downstream to the gene *KCNJ3*, a potassium channel gene that belongs to the cerebellum group and has a suggestive implication with epilepsy (Chioza et al.

2002). RNA-seq expression data available for one of the adult cerebellum samples included in this study suggests the presence of a 140 kb novel transcript located ca. 850 kb downstream to the gene *KCNJ3*. Interestingly, the genomic region comprising of the novel transcript and the last exon of *KCNJ3* was found to be deleted in two patients affected by developmental disorders with language delay and communication difficulties, for which a conclusive causal variant was not identified (Newbury et al. 2009).

## Discussion

In this study we generated a comprehensive atlas of transcription start sites for the human central nervous system, by sequencing at high depth (5 million reads per sample on average) CAGE libraries for 15 anatomically distinct regions of the CNS (Table 1). We identified 95,912 TCs, supporting the expression of 19,018 coding and non-coding genes, as annotated in Gencode v10. With this resolution, we broaden the landscape of brain gene expression: e.g. (Kang et al. 2011) reported the expression of 15'132 mainly coding genes in at least one region/developmental stage, while our previous study on a limited set of brain regions (Pardo et al. 2013) reported 16'888. It is likely that this increase is due to a combination of broader set of regions profiled, use of an annotation-independent profiling technique and extremely high sequencing depth. In particular we identified 4,779 intergenic TCs that represent bona fide novel transcripts. Notably 34.5% of them were supported by ChIP-seq H3K4me3 signature (marking sites of active transcription) derived from frontal lobe nuclei (Shulha et al. 2013): it is likely that with matched data this percentage would increase to 100%. Additionally an important fraction of the TCs for which we detect expression in brain (12.0%) maps to processed transcripts and retained introns, i.e. mainly representing non-coding transcripts associated to coding genes. This finding couples with recent publications (e.g. ENCODE Project Consortium 2012) demonstrating that pervasive transcription is a common feature of mammalian genomes. Although the functional meaning, and perhaps relevance, of this type of transcripts is under scrutiny and will require years of work to be

dissected, most likely at least part of them are functionally relevant, as suggested by the fact that 1. they contribute in setting apart CNS from other tissues and (Supplementary Figure 3) and 2. they show regionally biased patterns of expression that is comparable to that of coding genes. Additionally, these classes of poorly characterized transcripts contribute to the outstanding transcriptional complexity of CNS tissues (Supplementary Figure 2). The idea that brain structural and functional complexity is reflected at the transcriptional level was suggested in other studies before, such as (Jongeneel et al. 2005). Our data, based on a broader range of CNS and non-CNS samples supports this intriguing hypothesis. This is likely to reflect a combination of both the complex mixture of cell types in human nervous tissues (neurons, glial cells, blood vessels, microglial cells, macrophages etc.) and the rich set of novel brain specific transcripts, including those derived from repeat regions. The specificity and distinctiveness of CNS is clearly shown in our study (Figure 1, a-d), and it involves several layers of transcription: it's true at the level of coding genes, suggesting a structural component, it's true at the level of established and potential regulators (transcription factors and lncRNAs), and it holds true for the putative novel transcripts we identify. Interestingly on the same lines (Andersson et al. 2014) show that enhancer expression clearly group CNS tissues apart from other tissues. A feature of brain transcriptome that could be related to its complexity lies in expressed repeat regions, which have been linked to the evolution of gene expression and its regulation (Lynch et al. 2011). We show that TCs up-regulated in brain, in general slightly more often derived from repeats than TCs down-regulated in brain, are impressively enriched in simple and low complexity repeats and depleted of LTRs (Figure 2b). Similar observations were previously reported (Faulkner et al. 2009; Xu et al. 2010; Tyekucheva et al. 2011) and our study confirms that this is a specific feature of brain transcription. Additionally, we observed a significantly higher number of Alu (SINE family) and L1 (LINE family) repeats (Figure 2c) up-regulated in brain. This is particularly interesting as active somatic retrotransposition was described in the neural cell lineage and shown in particular to be a feature of mature human brain (Baillie et al. 2011). A class of non-coding transcripts that is gaining rising attention is the group of

lncRNAs. Very few of them have been characterized in detail so far but there is increasing evidence of their fundamental involvement in several biological processes, e.g. a lncRNA was recently described to regulate in mouse the translation of *UCLH1* (Carrieri et al. 2012). We identified over a thousand lncRNAs expressed in the CNS (1,407); in particular 419 were up-regulated in brain. This is an extremely valuable part of our dataset, completely unexplored and unknown, but potentially extremely relevant. Intriguingly, clustering of brain samples based on this subset of the data showed that these lncRNAs up-regulated in brain are expressed in a regionally biased manner (Figure 3e). The same observation applied for transcription factors (Figure 3f), raising the intriguing hypothesis that they coordinately regulate region-specific transcriptional programs. Interestingly several of the TFs up-regulated in the adult brain were known to have key roles during brain development, suggesting that some of the poorly annotated ones might have similar roles.

We were not able to identify expression signatures specific for every single region included in this study, probably because we are measuring average expression profiles in a very heterogeneous cellular population. This sounds particularly true in light of recent reports showing that mosaic copy number variants are a common feature in human (McConnell et al. 2013) and drosophila (Perrat et al. 2013) neuronal cell types. Future applications in brain transcriptomics will surely benefit of the recent advances in single-cell profiling. However, consistently with previous reports (Kang et al. 2011; Hawrylycz et al. 2012) we identified transcription profiles that tend to be similar between developmentally, functionally or morphologically related regions, with cerebellum showing the most distinctive expression signature and the largest number of differentially expressed TCs (Figure 4a). At the level of protein coding transcripts we find TCs mapping to known molecular markers for these regions, and we find considerable differences in expression of transcription factors, but similar to other analyses in this work, we want to highlight the large proportion (18.2%) of the differentially expressed TCs corresponding to non-coding transcripts.

Overall our findings clearly demonstrate the importance and power of using a completely unbiased approach to profile transcription. While projects such as



ENCODE have provided a wealth of new data on the transcriptional landscape of our genome, our finding that a substantial portion of transcription is brain specific and region specific clearly demonstrates the need to complement the available data with data directly derived from tissues. Moreover, the discovery that poorly annotated and non-coding transcripts significantly contribute to the specific transcriptional programs shows that functional annotation is still a long way to go. Our data therefore provides the scientific community with an important resource for interpreting the available genetic findings and provide testable hypotheses that can be followed up in the laboratory.

## **Materials and methods**

### ***Tissues, CAGE library preparation and sequencing.***

RNA was extracted from post mortem brain tissues obtained from the Netherlands Brain Bank (NBB; Amsterdam, The Netherlands). The donors were subjects without clinical signs of neurodegenerative or psychiatric disorders; age at death is reported in Supplementary Table 1. All brains were neuropathologically evaluated by an experienced neuropathologist and classified for neurofibrillary tangles (NFTs) stage 0-VI (Alafuzoff et al. 2008), amyloid-beta plaques score 0-C and Braak  $\alpha$ -synuclein stages 0-6 using the staging protocols of Brain Net Europe (BNE) and Braak (Alafuzoff et al. 2009, Alafuzoff et al. 2009, Braak et al. 2006). The dissection of all regions was performed from snap frozen human brain sections. Tissue was stored at -80°C until further processing. Total RNA was extracted and purified from tissues using the Trizol tissue kit according to instructions provided by the manufacturer (Invitrogen). RNA quality was assessed using the RNA Integrity Number (RIN) with the Agilent Total RNA Nano kit. All the samples used, their RIN values and library IDs are included in Supplementary Table 1. All the libraries were sequenced using the Heliscope single molecule sequencer. Library preparation and tag extraction were performed as described in (Forrest et al. 2014).

### ***Mapping, clustering and annotations.***

General data processing on the extracted tags was performed by the main project and is detailed in (Forrest et al. 2014). Briefly the steps include: 1. removal of the reads mapping to ribosomal RNA; 2. mapping of the remaining reads to the human genome (hg19 built) using the probabilistic mapping tool Delve; 3. removal of the tags with low mapping quality or mapping to the genome with less than 85% identity; 4. clustering of the CAGE tags into tag clusters (TCs) using the decomposition-based peak identification method (DPI. Kawaji et al. In preparation). Annotation files were built in the context of the main project and provided for each TC annotation information with respect to Gencode v10 gene model, CpG islands, TATA box and repeats from Repeat Masker in bed format. Expression tables containing tag counts and RLE (Relative Log Expression) normalized expression values were built by the main project (Forrest et al. 2014); when the expression “tpm” is used in the text, we refer to RLE normalized expression values. In this work a TC was considered to be expressed in brain if it counted at least 1 tpm in at least one of the brain libraries used; similarly a cluster was considered to be expressed in the FANTOM5 tissue collection if it counted at least 1tpm in at least one of the libraries in the collection.

### ***H3K4me3 ChIP-seq data from post mortem cortex.***

Data presented in (Shulha et al. 2013) was downloaded from <http://zlab.umassmed.edu/zlab/publications/ShulhaPLOSGen2013.html>. Only neuronal ChIP-seq samples with ages comparable to the adult donors in this study were included in the analysis (samples c25 to c31, average age 70 years). Peak calling was performed on each sample using MACS (Zhang et al. 2008) with parameters -bw=230 and -t=36 and using the input control available from (Shulha et al. 2013). A pool of 33,305 peaks was created considering all peaks called in at least one sample and merging adjacent peaks. Intersections were performed using windowBed (bedTools suite, Quinlan and Hall 2010), with a window size of 500 base pairs.

### ***Multidimensional scaling.***

Multidimensional Scaling (MDS), also known as principal coordinates analysis, is a data reduction technique similar to principal component analysis that can be used to visually represent the similarities (or differences) among a set of objects, in this case expression profiles. The MDS plots shown in this paper were performed using the R function *cmdscale()*. The matrix of pairwise euclidean distances between expression profiles used as input for the MDS plots was created with the R function *dist()* on log-transformed normalized data.

### ***Identification of TCs up-regulated in brain.***

To identify TCs up-regulated in brain we used the Bioconductor package edgeR (Robinson et al. 2010), dividing the samples of the FANTOM5 tissue collection (Supplementary Table 1) into two groups, that we named "brain" and "other". TCs satisfying the criteria "FDR < 1% after Benjamini-Hochberg correction and  $|\log FC| > 2$ " were considered differentially expressed between brain and other tissues. The TFs annotations used are provided in the as supplementary material of (Forrest et al. 2014). Expression heatmaps were performed using the R package gplots.

### ***Functional annotation.***

GO Biological Process and KEGG enrichment analysis on the genes up-regulated in brain was performed with DAVID (Sherman et al. 2007), using as background all the genes expressed in the reference set of brain samples. All the enrichments with FDR < 1% are reported in Supplementary Table 3.

### ***Identification of TCs differentially expressed across regions and shared patterns of differential expression.***

To identify TCs differentially expressed across regions we divided the CNS libraries into 15 groups (one for each of the brain regions under analysis) and used edgeR to perform the differential expression analysis (significance threshold: p-value < 0.01% after Benjamini-Hochberg correction and  $|\log FC| > 2$ ). We then combined the results of the pairwise comparisons requiring p-value < 0.01 after additional Bonferroni correction to adjust for multiple testing. To identify patterns of

expression shared between different brain regions we defined a matrix of “region-specificity” scores defined as “ $\log(\text{average tpm in region}) - \log(\text{average tpm across all brain samples})$ ”. Finally, we used the R (<http://www.r-project.org/>) implementation of the clustering algorithm k-means to partition the matrix into four groups that capture the major co-expression patterns observed (Supplementary Figure 5). These scores were used to create the heatmap in Figure 4b, performed using the R package gplots.

### ***Data Accessibility.***

All the data used in this paper is accessible through ZENBU, a fast, user-friendly and highly customizable genome browser (Severin et al. Submitted). Cell type and co-expression cluster specific annotations, motifs and transcription factors can be explored through the FANTOM5 Resource Browser (Shimoji et al. In preparation) and the FANTOM5 main portal available at <http://fantom.gsc.riken.jp/5/top/> (note to reviewers: the portal will be freely available after the main FANTOM5 paper is published). These resources and related references are described in more detail in (Forrest et al. 2014).

### ***Acknowledgements.***

We thank The Netherlands Brain Bank for providing the samples used in this study. This work was supported by the Neuroscience Campus Amsterdam, the Hersenstichting Nederland (PH and JSS), the Portuguese Foundation for Science and Technology (MF), the International Program Associate stipend from RIKEN (MV), The Frankopani Fund Scholarship (MV). FANTOM5 was made possible by a Research Grant for RIKEN Omics Science Center from MEXT to Yoshihide Hayashizaki and a Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan to YH. We would like to thank all members of the FANTOM5 consortium for contributing to generation of samples and analysis of the data-set and thank GeNAS for data production.

***Author contribution statements***

MF and MV did the analyses; MF, MV and PH wrote the manuscript, PR selected all samples, evaluated medical and pathological records and isolated RNA, JSS curated the list of disease loci, RA and AS identified active enhancers and analyzed their usage in brain samples, MI, NK, JK were responsible for data production, TL was responsible for tag mapping, HK managed the data handling, PC, HS, COD, YH and AF were responsible for FANTOM5 management and concept, AF, PC and PH designed the study.

***Disclosure declaration***

The authors declare no competing interests.

**Figure 1. Transcriptional specificity and complexity of the human brain transcriptome.**

**a.-d.** Multidimensional scaling representation of all TCs expressed in the FANTOM5 tissue collection for four subsets of the data, representing coding genes (a), transcription factors (b), lncRNAs (c) and putative novel transcripts (d). Brain samples (in grey) clearly separate from other tissues (in black) based on the expression profiles of all the subsets. The 9 samples that deviate from the tissue cluster in the direction of brain in (a) are: cerebrospinal fluid, vitreous humor, adipose tissue (donors 1 to 4), fetal eye, adult testis and adult retina. The 7 samples closest to brain at the level of TFs expression (b) are adult retina, fetal eye, adult testis and adipose tissue (donors 1 to 4). The two samples that deviate from both the CNS and the other tissues based on lncRNA expression are adult testis. **e.** Cumulative distribution of tags accounted for by the top 10,000 most highly expressed TCs for all the tissues in the FANTOM5 collection. The curves representing brain samples, in red, group in the lower part, indicating higher transcriptional complexity. **f.** Comparison of N50 values (number of TCs required to cover at least 50% of the library, providing a quantification of the curves in e.) counts for brain and other tissues, showing that brain has significantly higher transcriptional complexity with respect to other tissues.

**Figure 2. The genomic context of TCs up-regulated in brain.**

**a.** 36% of the TCs expressed across the FANTOM5 tissue collection are differentially expressed between brain and other tissues; in particular 18,626 TCs are up-regulated in brain. **b.** Fraction of TCs up- and down-regulated in brain (represented in red and blue respectively) mapping to a range of repetitive elements: TCs up-regulated in brain are clearly enriched in expressed simple and low complexity repeats, while they are depleted in LTRs. **c.** Relative number of expressed Alu and L1 repeats over the total number of expressed SINE and LINE respectively: TCs up-regulated in brain tend to be more frequently Alu- or L1-derived with respect to TCs down-regulated in brain.

**Figure 3. Up-regulated TCs in brain map to genes involved in brain function and to TFs and lncRNAs that are expressed in region-specific patterns.**

**a.** 10 most significantly enriched GO Biological Process terms in the set of genes up-regulated in brain: on the left side are reported the GO terms, the length of each bar represents significance (as  $-\log(\text{FDR})$ ), the numbers at the end of the bars represents the number of genes up-regulated in brain that map to the corresponding GO term. **b.** Similar representation as **a.** performed for KEGG pathways. The full list of enrichments are provided as supplementary material. **c.** The TF up-regulated in brain with highest expression is the transcriptional repressor *TSC22D4*. It's interesting to note that brain and other tissues express distinct isoforms that differ in their exon content. **d.** The up-regulated lncRNA with highest expression in brain tissues is *AC073479.1*, located in the chromosomal band 2p25.2. **e.** Heatmap representation of the expression profiles of the lncRNAs up-regulated in brain: each row represents a single TC and the corresponding expression, each column represents one sample. The colours of the top bar summarize the clustering of the brain regions into the four groups induced by the expression profiles of the TCs included in the heatmap (i.e. all TCs mapping to lncRNAs). **f.** Similarly as **e.** heatmap representation of the expression profiles of the TCs up-regulated in brain mapping to TFs.

**Figure 4. Differential expression across brain regions.**

**a.** Number of TCs identified as differentially expressed in each of the brain regions: cerebellum is characterized by the largest number of differentially expressed TCs (y axis indicates the number of differentially expressed TCs). Note that a cluster can be differentially expressed for more than one anatomical region. **b.** Heatmap representing the expression patterns of the TCs differentially expressed across regions. Each row represents a single TC and the corresponding expression on a per-region basis, showing that expression signatures are not characteristic of a single region but tend to be shared, suggesting a separation into four major groups of anatomical regions (top dendrogram). The four groups are summarized through the color bar below the dendrogram. **c.** Distribution of the biotypes of non-coding transcripts associated to CAGE TCs differentially expressed across brain regions.

The top represented classes are processed transcripts, retained introns and lncRNAs, accounting together for over 85% of all the differentially expressed clusters mapping to non-coding transcripts. **d.** Example of cerebellum specific intergenic cluster with RNA-seq support. This novel transcript is located ca. 850 kb downstream to the gene *KCNJ3*, over-expressed in cerebellum as well. The genomic region comprising of the novel transcript and the last exon of the gene *KCNJ3* was found to be deleted in two patients affected by developmental disorders with language delay and communication difficulties (see main text).

**Table 1. Regions of the CNS included in the study.**

Region name	Abbreviation used in the plots
Amygdala	amy
Caudate	caud
Cerebellum	cer
Globus pallidus	gIP
Hippocampus	hip
Locus coeruleus	locC
Medial Frontal Gyrus	MFG
Medial Temporal Gyrus	MTG
Medulla Oblongata	medO
Occipital Cortex	occC
Parietal Cortex	parC
Putamen	put
Spinal Cord	spC
Substantia Nigra	subN
Thalamus	thal



**Table 2: Summary of distribution of TCs.**

		Number
TCs	Expressed across all human tissues	122938
	Expressed in human brain	91643
	Brain-specific	26035
Coding Genes	Total number in all human tissues	93822 (76.3%)
	Expressed in human brain	72017 (78.6%)
	Brain-specific	19403 (74.5%)
Non-coding	Total number in all human tissues	21621 (17.6%)
	Expressed in human brain	15214 (16.6%)
	Brain-specific	4695 (18.0%)
Unannotated	Total number in all human tissues	7495 (6.1%)
	Expressed in human brain	4412 (4.8%)
	Brain specific	1937 (7.4%)

Figure 1

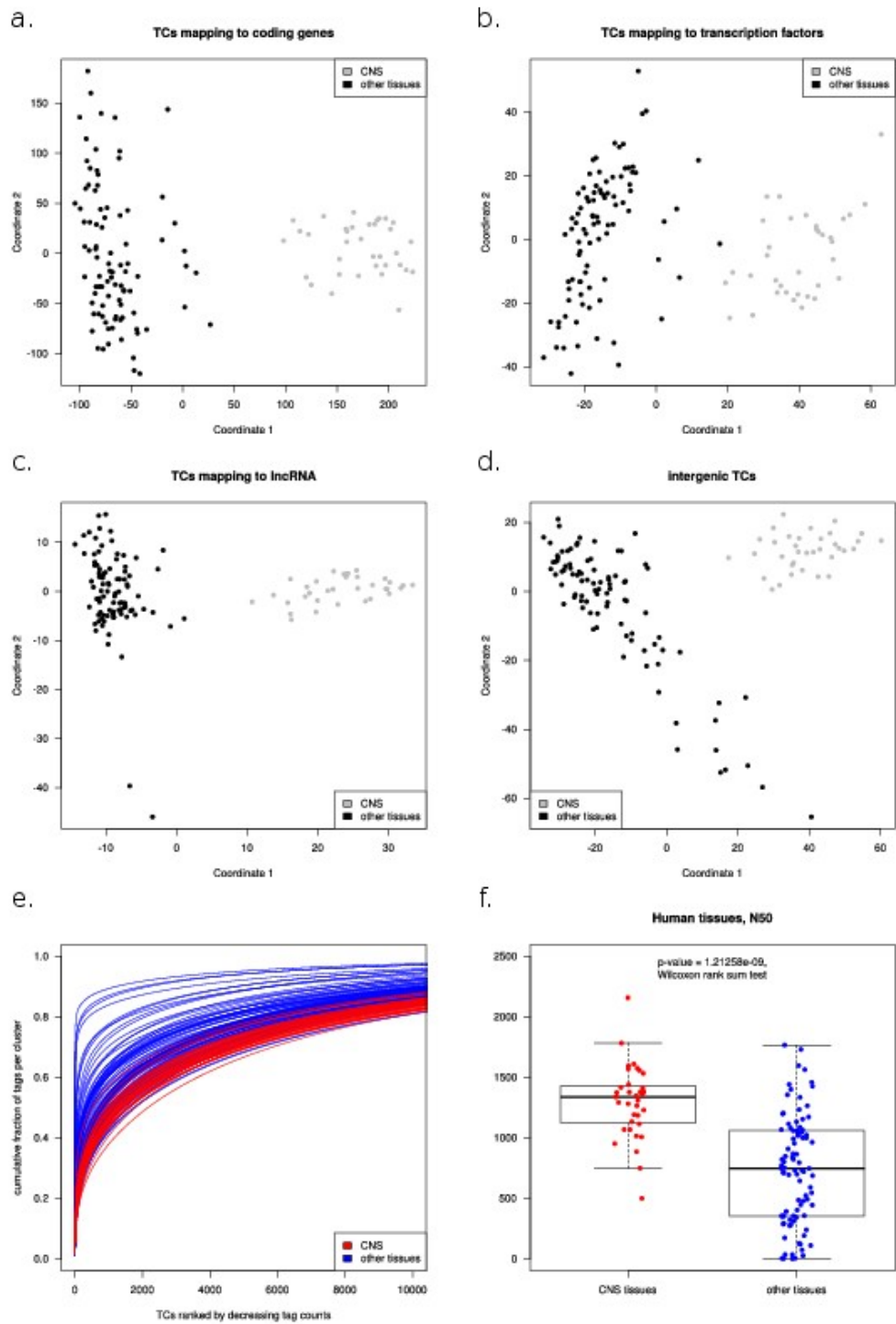
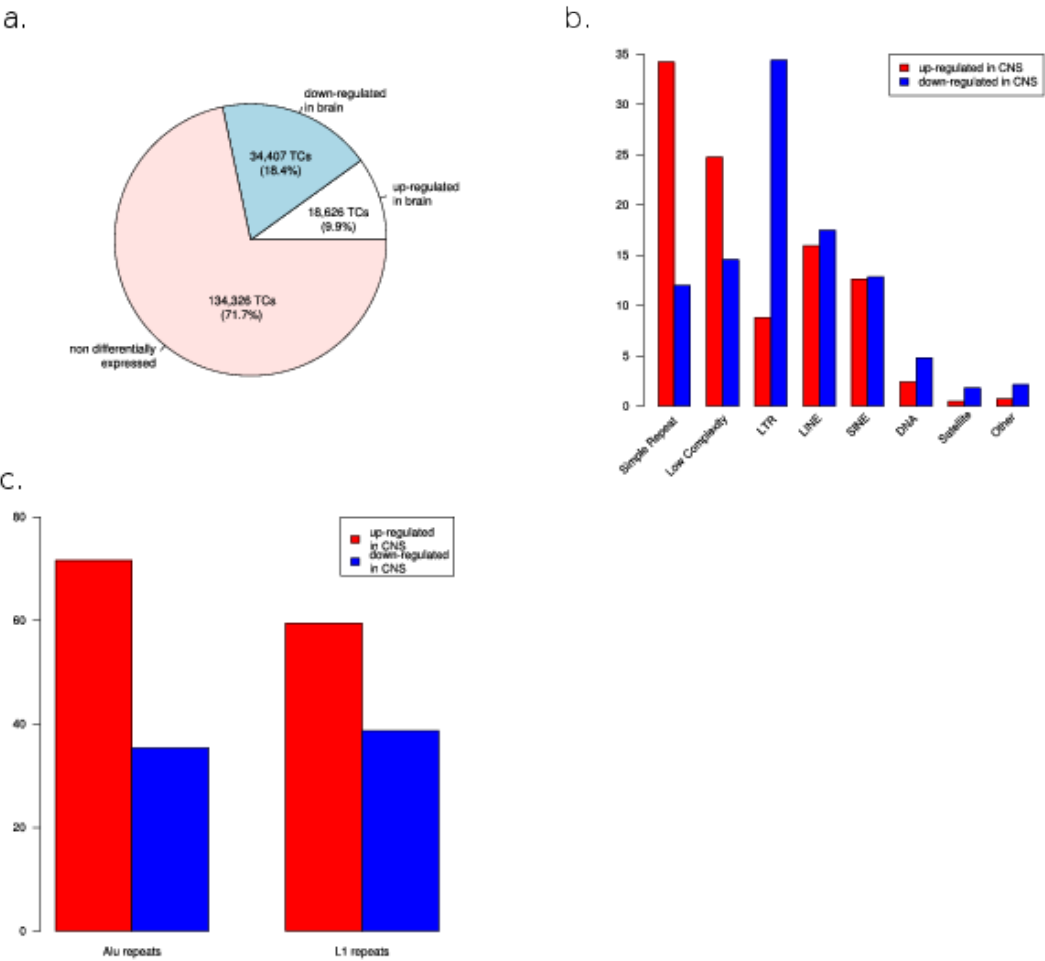
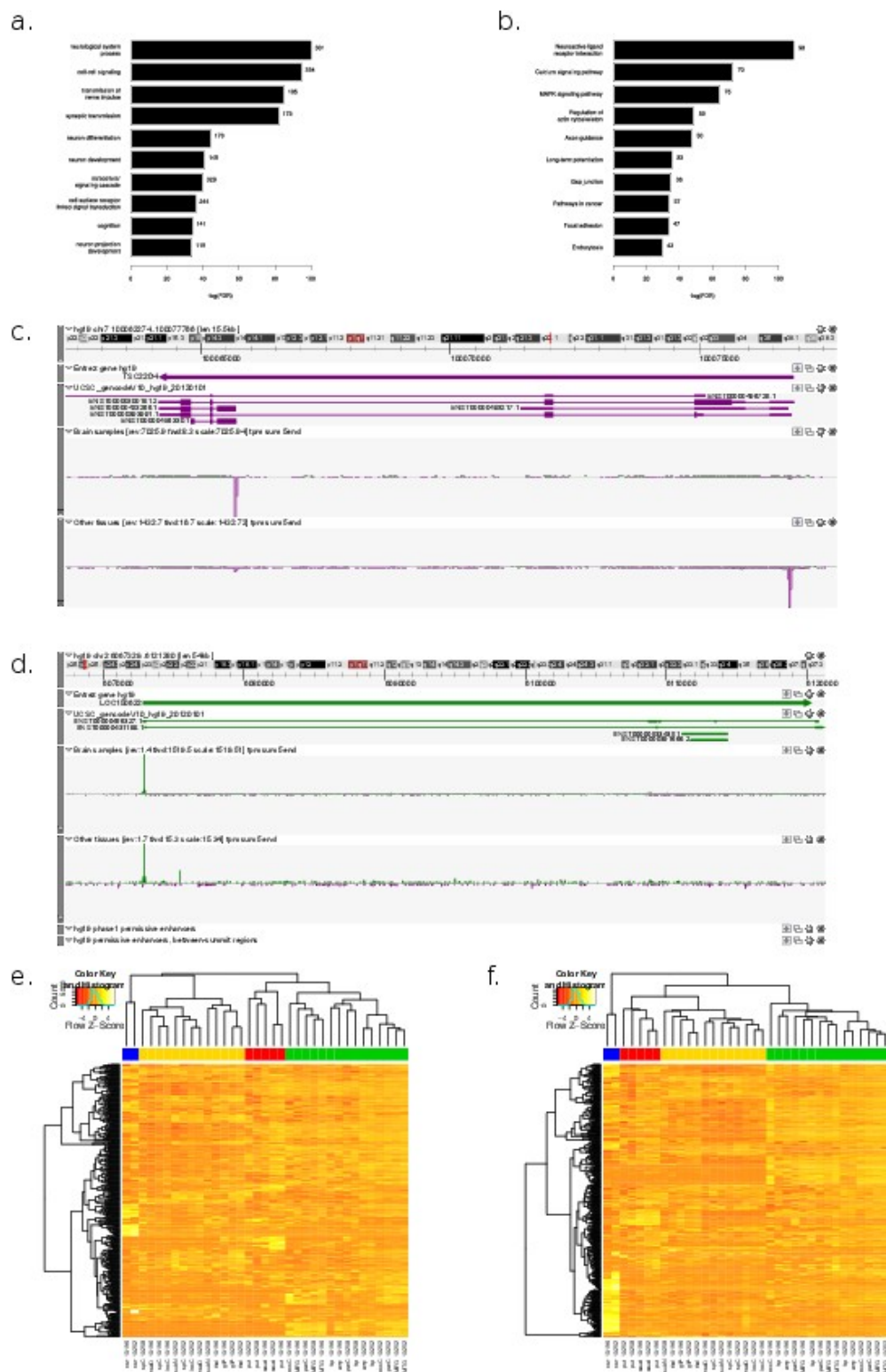


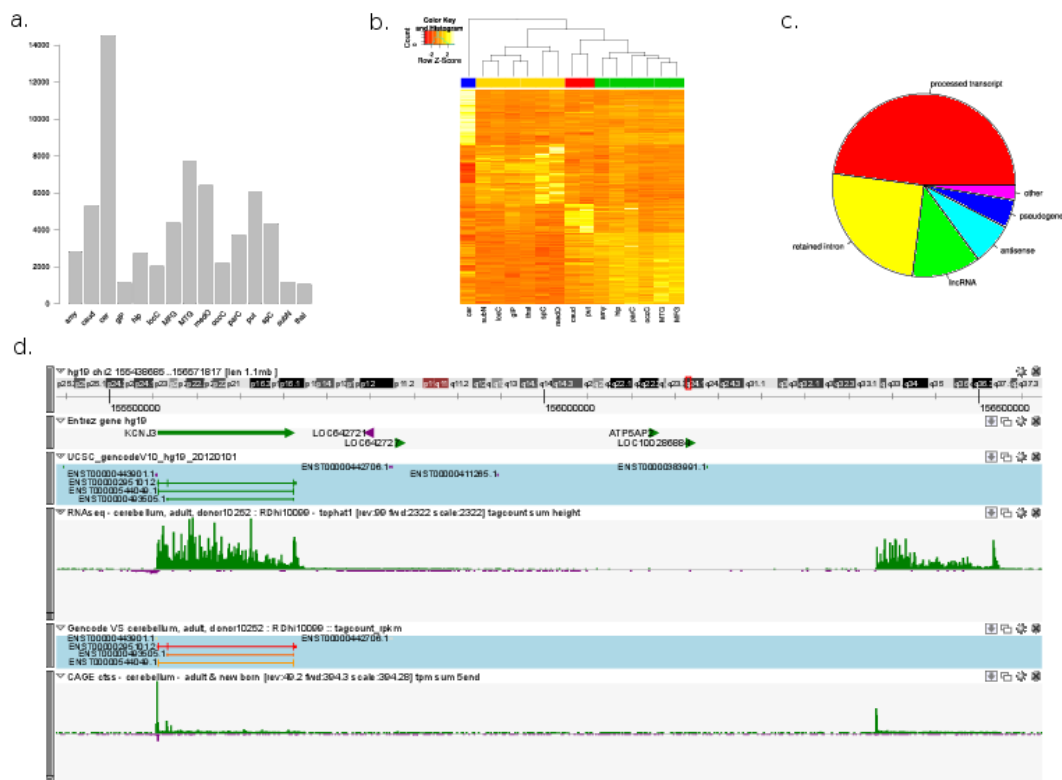
Figure 2



**Figure 3**



**Figure 4**



## References

- Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* 507(7493):455-61.
- Azevedo FA, Carvalho LR, Grinberg LT, Farfel JM, Ferretti RE, Leite RE, Jacob Filho W, Lent R, Herculano-Houzel S. 2009. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J Comp Neurol* **513**:532.
- Bacchelli E, Blasi F, Biondolillo M, Lamb JA, Bonora E, Barnby G, Parr J, Beyer KS, Klauck SM, Poustka A, et al. 2003. Screening of nine candidate genes for autism on chromosome 2q reveals rare nonsynonymous variants in the cAMP-

GEFII gene. *Mol Psychiatry* **8**: 916-924.

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, Talbot RT, Gustincich S, Freeman TC, Mattick JS, Hume DA, Heutink P, Carninci P, Jeddloh JA, Faulkner GJ. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**: 534.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**: 823-837.

Blank MC, Grinberg I, Aryee E, Laliberte C, Chizhikov VV, Henkelman RM, Millen KJ. 2011. Multiple developmental programs are altered by loss of *Zic1* and *Zic4* to cause Dandy-Walker malformation cerebellar pathogenesis. *Development* **138**: 1207.

Braak H, Bohl JR, Müller CM, Rüb U, de Vos RA, Del Tredici K. 2006. Stanley Fahn Lecture 2005: The staging procedure for the inclusion body pathology associated with sporadic Parkinson's disease reconsidered. *Mov Disord* **21**: 2042-2051.

Bulfone A, Smiga SM, Shimamura K, Peterson A, Puelles L, Rubenstein JL. 1995. T-brain-1: a homolog of Brachyury whose expression defines molecularly distinct domains within the cerebral cortex. *Neuron* **15**: 63-78.

Bulfone A, Wang F, Hevner R, Anderson S, Cutforth T, Chen S, Meneses J, Pedersen R, Axel R, Rubenstein JL. 1998. An olfactory sensory map develops in the absence of normal projection neurons or GABAergic interneurons. *Neuron* **21**: 1273-1282.

Canterini S, Bosco A, Carletti V, Fuso A, Curci A, Mangia F, Fiorenza MT. 2012. Subcellular TSC22D4 localization in cerebellum granule neurons of the mouse depends on development and differentiation. *Cerebellum* **11**:28.

Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, Pesce E, Ferrer I, Collavin L, Santoro C, et al. 2012. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* **491**: 454-457.

Chen CM, Wang HY, You LR, Shang RL, Liu FC. 2010. Expression analysis of an evolutionarily conserved metallophosphodiesterase gene, *Mpped1*, in the normal

and beta-catenin-deficient malformed dorsal telencephalon. *Dev Dyn* **239**: 1797-1806.

Chioza B, Osei-Lah A, Wilkie H, Nashef L, McCormick D, Asherson P, Makoff AJ. 2002. Suggestive evidence for association of two potassium channel genes with different idiopathic generalised epilepsy syndromes. *Epilepsy Res* 52:107.

Colantuoni C, Lipska BK, Ye T, Hyde TM, Tao R, Leek JT, Colantuoni EA, Elkahoul AG, Herman MM, Weinberger DR, Kleinman JE. 2011. Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature* **478**: 519-523.

Eicher JD, Powers NR, Miller LL, Akshoomoff N, Amaral DG, Bloss CS, Libiger O, Schork NJ, Darst BF, Casey BJ, Chang L, Ernst T, Frazier J, Kaufmann WE, Keating B, Kenet T, Kennedy D, Mostofsky S, Murray SS, Sowell ER, Bartsch H, Kuperman JM, Brown TT, Hagler DJ Jr, Dale AM, Jernigan TL, St Pourcain B, Davey Smith G, Ring SM, Gruen JR; Pediatric Imaging, Neurocognition, and Genetics Study. 2013. Genome-wide association study of shared components of reading disability and language impairment. *Genes Brain Behav* 12:792.

ENCODE Project Consortium, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

FANTOM Consortium, Suzuki H, Forrest AR, van Nimwegen E, Daub CO, Balwiercz PJ, Irvine KM, Lassmann T, Ravasi T, Hasegawa Y, et al. 2009. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet* **41**: 553-562.

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, Waki K, Hornig N, Arakawa T, Takahashi H, Kawai J, Forrest AR, Suzuki H, Hayashizaki Y, Hume DA, Orlando V, Grimmond SM, Carninci P. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41**: 563.

Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, et al. 2014. FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* 507(7493):462-70.

Fort A, Hashimoto K, Yamada D, Salimullah M, Keya CA, Saxena A, Bonetti A, Voineagu I, Bertin N, Kratz A, et al. Deep transcriptome profiling reveals that retrotransposons regulate pluripotency. Submitted.

Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, Menzel C, Chen W, Li Y, Zeng R, Khaitovich P. 2009. Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* **10**: 161. doi: 10.1186/1471-2164-10-161.

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L, et al. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**: 295-300.

Han W, Kwan KY, Shim S, Lam MM, Shin Y, Xu X, Zhu Y, Li M, Sestan N. 2011. TBR1 directly represses Fezf2 to control the laminar origin and development of the corticospinal tract. *Proc Natl Acad Sci U S A*. **108**: 3041-3046.

Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, van de Lagemaat LN, Smith KA, Ebbert A, Riley ZL, et al. 2012. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**: 391-399.

Hevner RF, Shi L, Justice N, Hsueh Y, Sheng M, Smiga S, Bulfone A, Goffinet AM, Campagnoni AT, Rubenstein JL. 2001. Tbr1 regulates differentiation of the preplate and layer 6. *Neuron* **29**: 353-366.

Hevner RF, Neogi T, Englund C, Daza RA, Fink A. 2003. Cajal-Retzius cells in the mouse: transcription factors, neurotransmitters, and birthdays suggest a pallial origin. *Brain Res Dev Brain Res* **141**: 39-53.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* **106**: 9362-9367.

Höglinger GU, Melhem NM, Dickson DW, Sleiman PM, Wang LS, Klei L, Rademakers R, de Silva R, Litvan I, Riley DE, et al. 2011. Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. *Nat Genet* **43**: 699-705.

Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, Yu C, Booth BW, Zhang D, Wan KH, et al. 2011. Genome-wide analysis of promoter



architecture in *Drosophila melanogaster*. *Genome Res* **21**: 182-192.

Hua P, Liu W, Kuo SH, Zhao Y, Chen L, Zhang N, Wang C, Guo S, Wang L, Xiao H, Kwan JY, Wu T. 2012. Association of Tef polymorphism with depression in Parkinson disease. *Mov Disord* 27:1694.

Hua P, Liu W, Zhao Y, Ding H, Wang L, Xiao H. 2012. Tef polymorphism is associated with sleep disturbances in patients with Parkinson's disease. *Sleep Med* 13:297.

Hutton M, Lendon CL, Rizzu P, Baker M, Froelich S, Houlden H, Pickering-Brown S, Chakraverty S, Isaacs A, Grover A, et al. 1998. Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature* **393**: 702-705.

Jones DC, Wein MN, Oukka M, Hofstaetter JG, Glimcher MJ, Glimcher LH. 2006. Regulation of adult bone mass by the zinc finger adapter protein Schnurri-3. *Science* 312:1223.

Joshi PS, Molyneaux BJ, Feng L, Xie X, Macklis JD, Gan L. 2008. Bhlhb5 regulates the postmitotic acquisition of area identities in layers II-V of the developing neocortex. *Neuron* **60**:258.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**: 182-187.

Jongeneel CV, Delorenzi M, Iseli C, Zhou D, Haudenschield CD, Khrebtukova I, Kuznetsov D, Stevenson BJ, Strausberg RL, Simpson AJ, Vasicek TJ. 2005. An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res* **15**: 1007-1014.

Kanamori-Katayama M, Itoh M, Kawaji H, Lassmann T, Katayama S, Kojima M, Bertin N, Kaiho A, Ninomiya N, Daub CO, Carninci P, et al. 2011. Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res* **21**: 1150-1159.

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AM, Pletikos M, Meyer KA, Sedmak G, et al. 2011. Spatio-temporal transcriptome of the human brain. *Nature* **478**: 483-489.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ.

2005. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**: D493-496.

Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, et al. 2006. CAGE: cap analysis of gene expression. *Nat Methods* **3**: 211-222.

Lenhard B, Sandelin A, Carninci P. 2012. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* **13**: 233-245.

Li YJ, Deng J, Mayhew GM, Grimsley JW, Huo X, Vance JM. 2007. Investigation of the PARK10 gene in Parkinson disease. *Ann Hum Genet.* 71:639.

Lynch VJ, Leclerc RD, May G, Wagner GP. 2011. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet* 43:1154.

Loewer S, Cabili MN, Guttman M, Loh YH, Thomas K, Park IH, Garber M, Curran M, Onder T, Agarwal S, et al. 2010. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet* **42**: 1113-1117.

Martínez-López MJ, Alcántara S, Mascaró C, Pérez-Brangulí F, Ruiz-Lozano P, Maes T, Soriano E, Buesa C. 2005. Mouse Neuron navigator 1, a novel microtubule-associated protein involved in neuronal migration. *Mol Cell Neurosci* **28**: 599-612.

Mattar P, Langevin LM, Markham K, Klenin N, Shivji S, Zinyk D, Schuurmans C. 2008. Basic helix-loop-helix transcription factors cooperate to specify a cortical projection neuron identity. *Mol Cell Biol* **28**: 1456-1469.

McConnell MJ, Lindberg MR, Brennand KJ, Piper JC, Voet T, Cowing-Zitron C, Shumilina S, Lasken RS, Vermeesch JR, Hall IM, Gage FH. 2013. Mosaic copy number variation in human neurons. *Science* **342**:632.

Newbury DF, Warburton PC, Wilson N, Bacchelli E, Carone S, Lamb JA, Maestrini E, Volpi EV, Mohammed S, Baird G, Monaco AP, IMGSAC. 2009. Mapping of partially overlapping de novo deletions across an autism susceptibility region (AUTS5) in two unrelated individuals affected by developmental delays with communication impairment. *Am J Med Genet Part A* **149A**: 588-597.

Pardo LM, Rizzu P, Francescato M, Vitezic M, Leday GG, Sanchez JS, Khamis A, Takahashi H, van de Berg WD, Medvedeva YA, van de Wiel MA, Daub CO, Carninci P, Heutink P. 2013. Regional differences in gene expression and promoter usage in aged human brains. *Neurobiol Aging* **34**: 1825-36.

Perrat PN, DasGupta S, Wang J, Theurkauf W, Weng Z, Rosbash M, Waddell S. 2013. Transposition-driven genomic heterogeneity in the *Drosophila* brain. *Science* **340**:91.

Plessy C, Pascarella G, Bertin N, Akalin A, Carrieri C, Vassalli A, Lazarevic D, Severin J, Vlachouli C, Simone R, et al. 2012. Promoter architecture of mouse olfactory receptor genes. *Genome Res* **22**: 486-497.

Quinlan AR and Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

R Core Team. 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Remedios R, Huilgol D, Saha B, Hari P, Bhatnagar L, Kowalczyk T, Hevner RF, Suda Y, Aizawa S, Ohshima T, et al. 2007. A stream of cells migrating from the caudal telencephalon reveals a link between the amygdala and neocortex. *Nat Neurosci* **10**: 1141-1150.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139-140.

Roider HG, Lenhard B, Kanhere A, Haas SA, Vingron M. 2009. CpG-depleted promoters harbor tissue-specific transcription factor binding signals--implications for motif overrepresentation analyses. *Nucleic Acids Res.* **37**: 6305.

Roth RB, Hevezi P, Lee J, Willhite D, Lechner SM, Foster AC, Zlotnik A. 2006. Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics* **7**: 67-80.

Saito T, Hanai S, Takashima S, Nakagawa E, Okazaki S, Inoue T, Miyata R, Hoshino K, Akashi T, Sasaki M, et al. 2011. Neocortical layer formation of human developing brains and lissencephalies: consideration of layer-specific marker expression. *Cereb Cortex* **21**: 588-596.

Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium,

Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, Holmans PA, Lin DY, Duan J, Ophoff RA, et al. 2011. Genome-wide association study identified five new schizophrenia loci. *Nat Genet* **43**: 969-976.

Severin J, Lizio M, Harshbarger J, Kawaji H, Daub CO, Hayashizaki Y, Bertin N, Forrest ARR. ZENBU: secured scientific collaborations, data integration and omics visualization. Submitted.

Sherman BT, Huang da W, Tan Q, Guo Y, Bour S, Liu D, Stephens R, Baseler MW, Lane HC, Lempicki RA. 2007. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics* **8**: 426. doi:10.1186/1471-2105-8-426.

Shulha HP, Cheung I, Guo Y, Akbarian S, Weng Z. 2013. Coordinated Cell Type–Specific Epigenetic Remodeling in Prefrontal Cortex Begins before Birth and Continues into Early Adulthood. *PLoS Genet* **9**: e1003433.

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguéz P, Doerks T, Stark M, Müller J, Bork P, et al. 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* **39**: 561-568.

Takahashi H, Lassmann T, Murata M, Carninci P. 2012. 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat Protoc* **7**: 542-561.

Talkowski ME, Maussion G, Crapper L, Rosenfeld JA, Blumenthal I, Hanscom C, Chiang C, Lindgren A, Pereira A, Ruderfer D, et al. 2012. Disruption of a Large Intergenic Noncoding RNA in Subjects with Neurodevelopmental Disabilities. *Am J Hum Genet* **91**: 1128-1134.

Turnbull J, Girard JM, Lohi H, Chan EM, Wang P, Tiberia E, Omer S, Ahmed M, Bennett C, Chakrabarty A, Tyagi A, Liu Y, Pencea N, Zhao X, Scherer SW, Ackerley CA, Minassian BA. 2012. Early-onset Lafora body disease. *Neuron* **135**:2684.

Tyekucheva S, Yolken RH, McCombie WR, Parla J, Kramer M, Wheelan SJ, Sabunciyan S. 2011. Establishing the baseline level of repetitive element expression in the human cortex. *BMC Genomics* **12**: 495.

Valen E, Pascarella G, Chalk A, Maeda N, Kojima M, Kawazu C, Murata M, Nishiyori H, Lazarevic D, Motti D, et al. 2009. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res* **19**: 255-265.

Wickham H. 2009. ggplot2: elegant graphics for data analysis. Springer, New York.

Xu AG, He L, Li Z, Xu Y, Li M, Fu X, Yan Z, Yuan Y, Menzel C, Li N, Somel M, Hu H, Chen W, Pääbo S, Khaitovich P. 2010. Intergenic and repeat transcription in human, chimpanzee and macaque brains measured by RNA-Seq. *PLoS Comput Biol*. **6**: e1000843.

Xu C, Aragam N, Li X, Villa EC, Wang L, Briones D, Petty L, Posada Y, Arana TB, Cruz G, Mao C, Camarillo C, Su BB, Escamilla MA, Wang K. 2013. BCL9 and C9orf5 are associated with negative symptoms in schizophrenia: meta-analysis of two genome-wide association studies. *PLoS One* 8:e51674.

Zeng H, Shen EH, Hohmann JG, Oh SW, Bernard A, Royall JJ, Glattfelder KJ, Sunkin SM, Morris JA, Guillozet-Bongaarts AL, et al. 2012. Large-Scale Cellular-Resolution Gene Profiling in Human Neocortex Reveals Species-Specific Molecular Signatures. *Cell* **149**: 483-496.

Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137. doi: 10.1186/gb-2008-9-9-r137.



# Chapter 5

## **Brain-specific noncoding RNAs are likely to originate in repeats and may play a role in up-regulating genes in cis**

Francescatto M, Vitezic M, Heutink P and Saxena A. 2014. Brain-specific noncoding RNAs are likely to originate in repeats and may play a role in up-regulating genes in cis. Accepted in Int. J. Biochem. Cell. Biol.





## **Abstract**

The mouse and human brain express a large number of noncoding RNAs (ncRNAs). Some of these are known to participate in neural progenitor cell fate determination, cell differentiation, neuronal and synaptic plasticity and transposable elements derived ncRNAs contribute to somatic variation. Dysregulation of specific long ncRNAs (lncRNAs) has been shown in neurodevelopmental and neuro-degenerative diseases thus highlighting the importance of lncRNAs in brain function. Even though it is known that lncRNAs are expressed in cells at low levels in a tissue-specific manner, bioinformatics analyses of brain-specific ncRNAs has not been performed. We analyzed previously published custom microarray ncRNA expression data generated from twelve human tissues to identify tissue-specific ncRNAs. We find that among the 12 tissues studied, brain has the largest number of ncRNAs. Our analyses show that genes in the vicinity of brain-specific ncRNAs are significantly up regulated in the brain. Investigations of repeat representation show that brain-specific ncRNAs are significantly more likely to originate in repeat regions especially DNA/TcMar-Tigger compared with non-tissue-specific ncRNAs. We find SINE/Alus depleted from brain-specific dataset when compared with non-tissue-specific ncRNAs. Our data provide a bioinformatics comparison between brain-specific and non tissue-specific ncRNAs.

## **Introduction**

Long noncoding RNAs (lncRNAs) are generally regulatory in nature, modulating transcriptional silencing through chromatin modification<sup>2,3</sup>, transcriptional activation due to gene proximity<sup>4</sup> or enhancer function<sup>5,6</sup> or participating in up-regulating protein translation through inverted repeats sequences as in UCHL1<sup>1</sup>. Initially lncRNAs were defined as transcripts longer than 200 nucleotides that were incapable of coding for more than 100 amino acids [59]. This arbitrary definition was contradicted by the demonstration of functional polypeptides shorter than 100

amino acids <sup>7</sup>, the evidence that lncRNAs harbor open reading frames (ORFs) and the discovery of bi-functional RNAs which not only produce proteins but also function as regulatory lncRNAs <sup>8-12</sup>. Interrogation of whole transcriptomes of cells by the ENCODE consortium has revealed that even though 62% of the genomic bases are reproducibly represented in transcribed long RNA molecules or GENCODE exons, only 2.94% are represented in GENCODE annotated exons of protein coding genes <sup>13-15</sup>. A large proportion of the human genome appears to generate RNAs at low expression levels, 80% of lncRNAs transcripts and 25% of protein coding (pc) transcripts are detected at 1 copy (or less) per cell suggesting that the expression of nc transcripts may be limited to subpopulations of cells <sup>15</sup>. These nc transcripts are generally located within the nucleus, may be adenylated or polyadenylated and originate from intergenic as well as genic loci <sup>15, 16</sup>. Based on data generated in the ENCODE project, a GENCODE v7 catalogue has been created for human lncRNAs containing 14,880 non coding transcripts arising from 9,277 ncRNA genes <sup>16</sup>. Investigation of peptide signatures through MS / MS analyses revealed that 92% of lncRNAs listed in the GENCODE v7 catalogue have no protein coding competency <sup>17</sup>. While most lncRNAs are bioinformatically derived, their functional evaluation is difficult due to their low levels of expression and absence of defined landmarks such as sequence similarities.

Due to the fact that lncRNAs display a lower level of sequence conservation across species <sup>32, 33</sup>, investigations on lncRNA expression and function call for species-specific analyses. For example, the 2 kb long multifunctional lncRNA *HOTAIR* can potentially silence hundreds of gene targets via independent interactions with the PRC2 and LSD1 complexes <sup>34</sup> and is required for silencing genes within the HOXD cluster <sup>35</sup> and in determining the metastatic potential of breast <sup>36</sup> and nasopharyngeal cancers <sup>37</sup> in humans but in mouse, *hotair* is structurally different and displays no functional similarities to human *HOTAIR*<sup>38</sup>. Although RNA-seq has been largely instrumental in the discovery, assembly and annotation of the current catalogue of lncRNAs, due to the low expression levels of ncRNAs and the depth required to rule out false negatives, it is impractical to utilize sequencing to ascertain the presence of lncRNAs in tissues to generate tissue-specific ncRNA datasets [16; 39]. Further, due to strict size selection during

current library preparation protocols, it is difficult to conduct combined analyses of long and short RNAs using RNA-seq datasets [3; 16]. Overall microarray platforms are better suited for studies designed to detect the presence or absence of annotated long and short ncRNAs across tissues <sup>16, 39</sup>.

RNA-seq and custom lncRNA microarray analyses of multiple human organs suggest that lncRNAs display higher tissue-specificity compared with pc transcripts <sup>16</sup> [60]. This appears to be particularly true for the mouse brain where lncRNA expression is reported to be region-specific as shown in Mercer et al <sup>18</sup> who extracted 849 ncRNAs from the Allen Brain Atlas in-situ hybridization assays and found neuro-anatomical and cell subtype-specific expression in the adult mouse brain <sup>18</sup>. Many studies on lncRNAs and their function and dysregulation have focused solely on the brain and brain related disorders. Recent literature suggests that ncRNAs play a vital role in the brain [61] not only during development <sup>19 20</sup> [62; 63; 64], neural stem cell differentiation <sup>21</sup> and protein translation at synapses, but also contribute to higher order functions such as long-term memory formation <sup>22</sup> and synaptic plasticity <sup>20, 23 24, 25</sup>. Dysregulation of specific ncRNAs has been shown in neurodegenerative diseases with cognitive decline such as Alzheimer's disease, Parkinson's disease and Huntington's disease <sup>26-28, 29, 30</sup>. Further, short processed ncRNAs are also known to affect brain function and are expected to be dysregulated in diseases <sup>31</sup>. Two independent studies have identified lncRNAs expressed in the mouse <sup>18, 20</sup> and human brain <sup>18, 20</sup>, however it is not known if all these lncRNAs are functional or if they are expressed exclusively in the brain. Our aim in this study is to identify ncRNAs expressed solely in the brain and investigate them bioinformatically, to uncover distinctions from ncRNAs that are expressed in more than one tissue.

Nielsen et al designed 60 nucleotide custom microarray probes for 26,910 potentially functional ncRNAs of size greater than 60 nucleotides extracted from various databases <sup>39</sup>. They selected ncRNAs probes based on conservation, expression and active chromatin marks and determined the presence or absence of these ncRNAs in 12 human tissues <sup>39</sup>. After discarding probes that mapped to

mitochondrial RNAs, pseudogenes and those probes that multi-mapped to repeats, they identified a set of 12,115 ncRNAs of which 3,513 were expressed in at least 1 tissue above background. We analyzed this microarray dataset<sup>39</sup> to identify ncRNAs expressed in a single tissue. Our data show that among the 12 tissues studied, brain harbors the largest number of single tissue ncRNAs, which we call brain-specific ncRNAs. Further analyses of brain-specific ncRNAs reveal that they are longer than 200 nucleotides and map close to genes up regulated in the brain. Analyses of repeats in their transcript body reveal that brain-specific ncRNAs originate more often in repeats and are likely to be depleted in SINE/Alu.

### **Brain-specific ncRNAs are located in the vicinity of genes up regulated in the brain**

We analyzed the microarray dataset from Nielsen et al to identify ncRNAs whose expression was restricted to a single tissue and found that brain had the largest number of such ncRNAs (n=303) (Table 1, Figure 1a). We refer to this dataset as brain-specific ncRNAs. To verify that the microarray dataset used in our study captured ncRNAs missed by RNA-seq, we overlapped brain RNA-seq dataset from Wang et al (Nature 456 470-476) with our microarray based brain-specific ncRNAs dataset derived from Nielsen et al (ref). Our data show that out of 303 brain specific ncRNAs, only 96 were found in the RNA-seq dataset, an overlap of 32% (Supplementary Table 1 worksheet overlap\_microarray\_RNA-seq). For example we did not find ncRNAs in the intron of Brain specific angiogenesis inhibitor 3 (*BAI3*), Neuroligin 1 (*NLGN1*), as well as ncRNAs distal to genes such as potassium voltage-gated channel (*KCNH5*) and Myelodysplastic syndrome 2 translocation associated gene (*MDS2*) in the RNA-seq dataset.

Since ncRNA transcription can affect the expression of nearby genes positively or negatively<sup>4-6</sup>, we extracted the pc genes nearest to the brain-specific ncRNAs from GENCODE v17 (Supplementary Table 1 worksheet closest\_gene\_and\_distance), which provided a list of 283 unique genes. In view of

a recent publication suggesting a positive correlation between the expression of ncRNAs and mRNAs within 20 kbs of each other [16], we selected protein-coding transcripts within 20 kb, thus limiting our analyses to 187 protein coding genes in the vicinity of 303 brain-specific ncRNAs <sup>16</sup>. Expression analysis shows that 91 out of 187 pc genes (48.7%) were up-regulated in the brain (p-value <0.01, FDR 0.05) (Figure 1b, Supplementary Table 1 worksheet upTissue\_20-100kb). Next, we conducted functional enrichment analysis on the 187 pc genes using DAVID to identify if they shared any similarities. Investigation of the GO terms associated with these genes failed multiple testing thresholds (Supplementary Table 1 worksheet GO\_20kb). Extending our window to genes within 50 and 100 kb also failed to reveal brain-specific GO terms (Supplementary Table 1 worksheets GO\_50kb and GO\_100kb). We also conducted STRING network analyses <sup>40</sup> to investigate if the pc genes within 20 kb were functionally related to each other. Our data failed to reveal widespread functional interaction between the genes in the vicinity of brain-specific ncRNAs (Figure 1c).

### **Association of brain-specific ncRNAs with repeats**

Almost 45% of the human genome is made of repeat DNA sequences called transposable elements (TEs) or mobile elements <sup>41, 42</sup>. TEs called retrotransposons are capable of being transcribed into partial or full length RNAs and transported into the cytoplasm for translation of specific proteins. These proteins reverse transcribe the retrotransposon RNA into DNA and insert it back into the human genome <sup>42</sup>. The reverse transcription does not always extend to the end of the transcript resulting in partial, transpositionally incompetent copies inserted in the genome. Due to the potential mutagenicity caused by such amplification, transpositionally competent full-length repeat elements are heavily methylated in adult cells <sup>43</sup>, however the transpositionally incompetent copies are abundantly represented in the transcriptomes of human and mouse cells and suggested to have a role in transcriptional regulation <sup>44</sup>. Intersection of TE catalogs with lncRNA databases has shown that TEs are represented in 83.4% of lncRNAs and merely

39.1% of protein coding RNAs <sup>45</sup>. Recently, the retrotransposition competent TEs were shown to be active in differentiating neural progenitor cells <sup>46</sup> and in adult brain <sup>47, 48</sup> thus making a case for somatic insertion events in mouse and human brain. Among all tissue-specific ncRNAs identified in this study, brain had the highest number of specific ncRNAs originating within 100 bp of a repeat element followed closely by muscle (Supplementary Table 2\_Repeats\_at\_origin\_tissue\_wise). Comparison of brain-specific and non-tissue-specific ncRNAs (ncRNAs expressed in more than 1 tissue) showed that a significantly higher percentage of brain-specific ncRNAs originate from repeat regions (66.7% vs. 57.3%, p-value < 0.01) (Figure 2a, Table 2 and Supplementary Table 2, worksheet brain\_repeats\_at\_origin).

Since LTRs are reported as frequently found at the transcription start sites (TSSs) of long intergenic ncRNAs (lincRNAs) <sup>44, 45, 49, 50</sup> we investigated the representation of major repeat families at the TSSs of brain-specific and non-tissue-specific ncRNAs. Our data showed an abundance of LINE L1 and SINE Alu elements at the TSSs of ncRNAs (Figure 2b and Supplementary Table 2 worksheets brain\_repeats\_at\_origin and brain\_repeat\_names\_at\_origin). Further investigations revealed that brain-specific ncRNAs were significantly enriched in DNA/TcMar-Tigger at origin in comparison with non-tissue-specific ncRNAs (p < 0.05) (Table 2 and Supplementary Table 2 worksheet repeats\_at\_origin).

We also looked for raw number of repeats embedded within the body of brain-specific and non-tissue-specific ncRNAs. We found that 91% of brain-specific ncRNAs harbor repeats within the transcript body, a number marginally higher than the non-tissue-specific ncRNAs (85.1%) (p < 0.01) (Figure 2c, Table 2 and Supplementary Table 2, worksheet intersecting\_repeats\_transcript). Analyses of raw numbers of repeat families within transcript body showed that brain-specific ncRNAs were significantly depleted in SINE/Alu and SINE/MIR but enriched in LINE/L1, LTR/ERV1 and DNA repeats in comparison with non-tissue-specific ncRNAs (p < 0.001) (Figure 2d, Table 2 and Supplementary Table 2 intersecting\_repeats\_transcript). To exclude inaccuracies resulting from the length of transcripts in this analysis, we repeated this analysis to investigate base pair coverage by repeat families in both datasets. Our data show that despite

significantly different raw numbers, the overall percentage of repeat-derived sequence for brain-specific and non-tissue-specific ncRNAs is similar (49.8% vs. 48.3% respectively) (Figure 2E). Family wise distribution per base pair coverage shows that SINE/Alus were depleted and LINE/L1 were enriched in transcript body of brain-specific ncRNAs when compared with non-tissue-specific ncRNAs (Figure 2F, Supplementary Table 2 worksheet intersecting\_repeats\_bp\_cov) but the difference was not statistically significant.

### **Poor representation of noncoding Human accelerated regions in Brain-specific ncRNAs**

Certain regions of the human genome are termed highly accelerated regions (HARs) because they are found conserved in human and primate species only <sup>51-54</sup>. A total of 2,649 loci in humans have been designated noncoding HARs (ncHARs) and they were recently shown to function as enhancers <sup>55</sup>. We investigated if any of the brain-specific ncRNAs identified by us were arising from the ncHARs. Our data show minimal representation of ncHARs with only 5 brain-specific ncRNAs harboring ncHARs (Table 1). In general we find HAR regions not highly represented in our dataset of brain-specific ncRNAs (Table 1).

### **Discussion**

We derived a set of ncRNAs expressed exclusively in the brain from the ncRNA dataset published in Nielsen et al [39] and investigated them for characteristics of nearby genes, associations with repeat regions and representation of ncHARs. Nielsen et al presented a stratified assessment of tissue-specific ncRNAs divided by genomic location (intergenic, intronic and antisense) and analyzed all transcripts for expression, conservation and overlap with epigenetic marks [39]. In our manuscript we extracted ncRNAs expressed only in the brain and compared them to ncRNAs expressed in more than one tissue and investigated them with

respect to genes in the vicinity, repeat elements at origin and in transcript body and relationship to human accelerated regions. Our data show that brain expresses the highest number of tissue-specific ncRNAs. Surprisingly, among the 1,744 non-tissue-specific ncRNAs, only 109, were found expressed in all 12 tissues confirming earlier reports that ncRNAs are generally tissue-specific<sup>15, 16</sup>. Analyses of GO terms of genes within 20 kb of brain-specific ncRNAs reveals that even though these ncRNAs are expressed exclusively in the brain, they are not located in the vicinity of protein coding genes with known brain-specific functions. Extending our window of analyses to 50 and 100 kb did not yield additional functionally relevant pc genes. Further, string network analyses revealed that the proteins encoded by neighboring genes were functionally unrelated to each other suggesting that brain-specific ncRNAs did not regulate expression of brain-specific biological or functional networks in cells. This finding may reflect the fact that we limited our analyses to genes within 100 kb, which may be too restrictive. Nevertheless, our data reveal that brain-specific ncRNAs are located within 20 kb of genes that show significant enrichment in genes up-regulated in brain suggesting a cis regulatory role for brain-specific ncRNAs.

Even though the dataset used by us was filtered for repeat regions [39], we found that a significantly higher number of brain-specific ncRNAs originate near repeats in comparison with non-tissue-specific ncRNAs ( $p < 0.05$ ). In both datasets (brain-specific and non-tissue-specific ncRNAs) we find that a total of over 50% ncRNAs originate near LINE/L1 or SINE/Alu and less than 15% originate in LTR/ERVs.

Our data are in partial agreement with that from Kelley et al who showed significant enrichment of LTR/ERVs and depletion of LINE L1 and SINE/Alu at start sites of all lincRNAs<sup>45</sup> when compared with pc genes and genomic abundance. This discrepancy may result from the fact that there is poor co-relation between RNA-seq and microarray data for low expressed lincRNAs<sup>16</sup>, there were differences in data processing (repeats were filtered out of the microarray dataset and RNA-seq data were filtered to retain only those multi-mapping reads that mapped to less than 50 places with high confidence), differences in data analyses



(we restricted our analyses to 100 bp upstream of start sites and Kelley et al looked for repeats up to 2000 nt upstream) and differences in datasets (Kelley et al compared lincRNAs with pc RNAs and showed enrichment / depletion over genomic abundance while we compared brain-specific with non-tissue-specific lncRNAs). Given the non random nature of distribution of SINE/Alus in the genome (Lander Nature 2001, Grover 2004 Bioinformatics, Medstrand Cytogenet Genome Research 2005) and the fact that SINE/Alu may offer binding sites to development transcription factors (Oei 2004 Genomics, Polak 2006 BMC Genomics), this finding recommends further investigation into the expression of brain-specific ncRNAs during development.

Investigation of the percentage of ncRNAs harboring repeats shows that even though a significantly higher number of brain-specific ncRNAs intersect with repeats when compared with non-tissue-specific ncRNAs, brain-specific ncRNAs are significantly depleted in SINE /Alu and SINE / MIR and marginally enriched in LINE/L1 and LTR/ERV. The depletion of SINE/Alu from brain-specific ncRNA dataset may just reflect the tissue-specificity of this dataset since lncRNAs that contain SINE/Alu are expected to be less tissue-specific (Kelley). In view of the fact that TE containing lncRNAs are poorly expressed and that the presence of SINE/Alu confers greater expression (Kelley), our data indicate that brain-specific ncRNAs, which are rich in repeats and depleted in SINE/Alu, may be poorly expressed. Analyses by base pair coverage of ncRNAs by repeat families indicates that even though brain-specific ncRNAs are depleted in SINE/Alu and marginally enriched in LINE/L1, the difference was not statistically significant.

We found minimal representation of ncHARs in tissue-specific ncRNAs, which may be due to the fact that ncHARs are generally less than 200 nucleotides in length <sup>53</sup>, <sup>55</sup> and therefore may have been excluded from the custom microarray probe design due to length constraints.

Our work presents the first analyses of brain-specific ncRNAs in the context of neighboring genes and repeat elements. It is tempting to speculate that the brain-

specific ncRNAs play a role in up-regulating the expression of genes in their vicinity. Overall our data show that brain-specific ncRNAs are significantly likely to be poorly expressed and originate in repeat regions especially DNA/TcMar-Tigger. Further tissue-specific analyses of larger datasets and experimental validation will be important to decipher the role of brain-specific ncRNAs and the importance of repeats in the brain transcriptome.

**Table1 : Summary of tissue-specific ncRNAs**

Number of tissue -specific ncRNAs detected in the custom microarray dataset (Nielsen et al). Percentages out of total are indicated in parenthesis

Tissue	Number of tissue-specific ncRNAs (percentage out of total 3515 ncRNAs)	Number of tissue-specific ncRNAs harboring repeat elements (percentage out of tissue-specific ncRNAs in that tissue)	Number of tissue-specific ncRNAs intersecting with ncHARs (percentage out of tissue-specific ncRNAs in that tissue)
Bladder	95 (2.7)	49 (51.58)	1 (1.05)
Brain	303 (8.6)	202 (66.67)	5 (1.65)
Breast	104 (3.0)	64 (61.54)	2 (1.92)
Colon	132 (3.8)	79 (59.85)	2 (1.52)
Heart	150 (4.3)	92 (61.33)	5 (3.33)
Kidney	69 (2.0)	41 (59.42)	2 (2.90)
Liver	128 (3.6)	76 (59.38)	5 (3.91)
Lung	177 (5.0)	107 (60.45)	8 (4.52)
Muscle	83 (2.4)	55 (66.27)	2 (2.41)
Ovary	169 (4.8)	101 (59.76)	4 (2.37)
Prostate	247 (7.0)	144 (58.30)	4 (1.62)
Skin	112 (3.2)	64 (57.14)	1 (0.89)

**Table 2: Repeat analyses of ncRNAs**

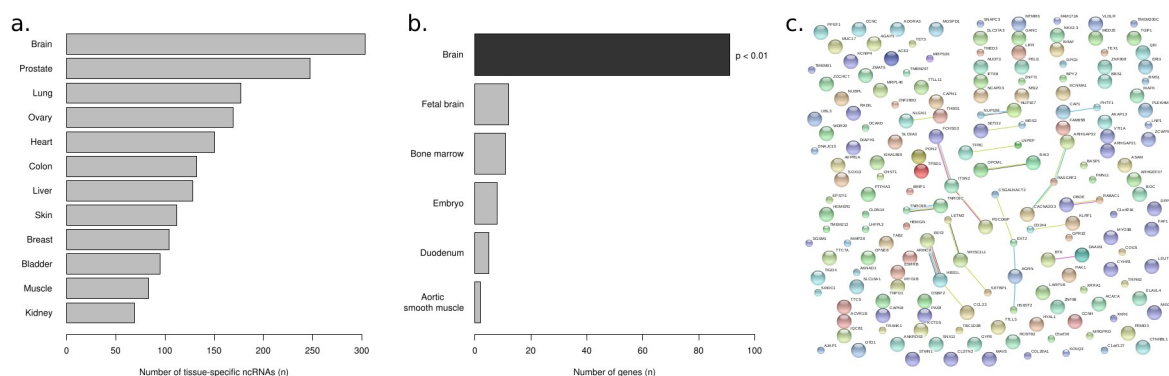
Differences in repeat regions represented in brain-specific ncRNAs, other-tissue-specific ncRNAs and non-tissue-specific ncRNAs with respect to numbers and location. See Supplementary Table 2 for a full list. Bold text indicates statistically significant difference.

<b>P Values of significance, Fishers test</b>	
<b>Repeat comparison</b>	<b>Brain-specific vs non-tissue-specific</b>
Number of repeats at origin	<b>0.002</b>
LINE L1/at origin	0.56
SINE Alu at origin	0.21
SINE MIR at origin	0.18
LTR / ERVL at origin	0.68
Number of repeats in transcript body	<b>0.0029</b>
LINE L1 in transcript body by count	<b>0.0007</b>
SINE Alu in transcript body by count	<b>4.2e-20</b>
SINE MIR in transcript body by count	<b>3.7e-08</b>
LTR/ERVL in transcript body by count	<b>0.003</b>

## Figure 1. Brain harbors the largest number of tissue-specific ncRNAs

**a.** Each bar represents the number of tissue-specific ncRNAs identified in corresponding tissues as labeled. Brain shows the highest number of tissue-specific ncRNAs (303, representing 17.1% of all tissue-specific ncRNAs), followed by prostate with 247. The counts for all tissues are shown in Table 1. **b.** Panel showing DAVID "UP\_TISSUE" category. The protein coding genes located within 20kb of brain-specific ncRNAs are enriched for genes up-regulated in brain (Benjamini-Hochberg adjusted  $p$ -value  $< 0.01$ ). The length of each bar corresponds to the number of genes present in each group. **c.** STRING protein-protein interaction network on the protein coding genes located within 20kb of brain-specific ncRNAs: each circle (node) represents a protein-coding gene; nodes are connected when evidence exists that they interact at the protein level. The presence of very few connected nodes suggests that most of the protein-coding genes located within 20kb of brain-specific ncRNAs are not currently known to interact functionally.

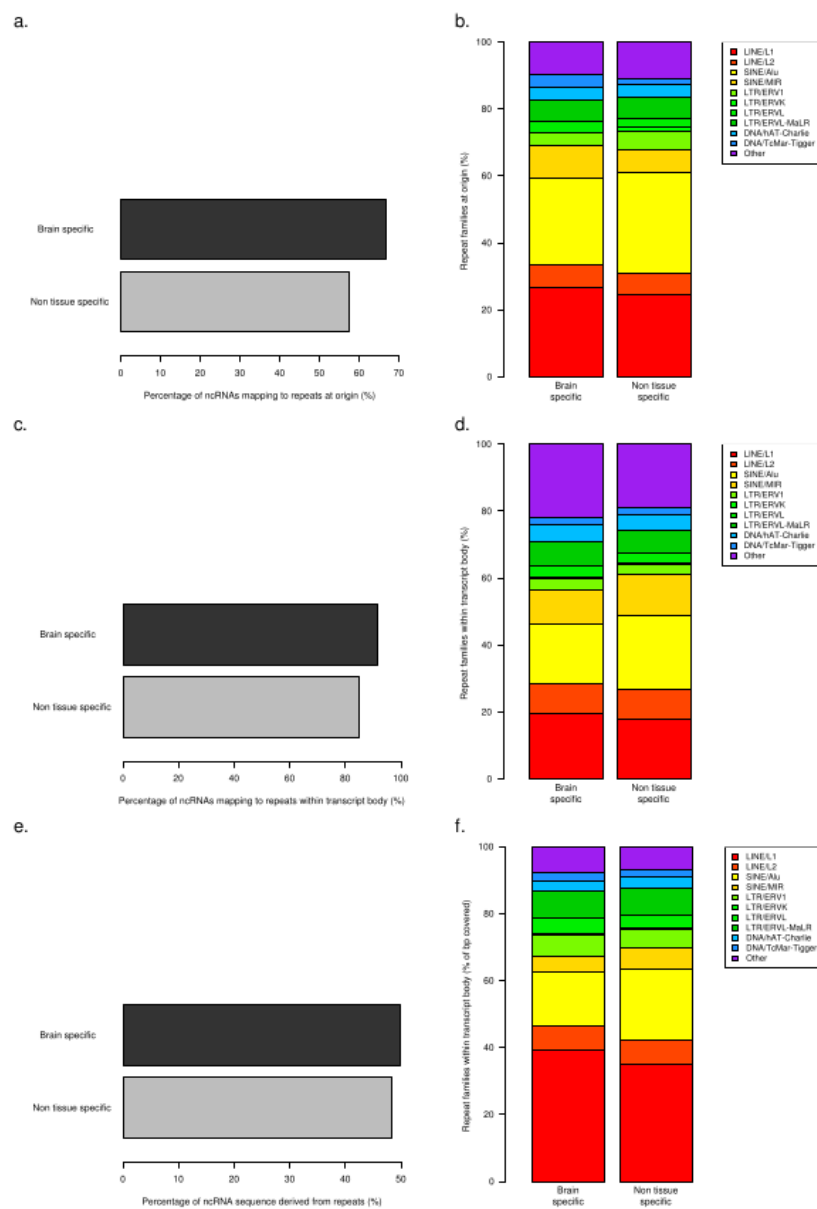
**Figure 1**



**Figure 2. Repeat elements composition of expressed brain-specific and non-tissue-specific ncRNAs.**

**a.** Percentage of brain-specific and non-tissue-specific ncRNAs originating within 100bp of repetitive elements. Overall a large percentage of ncRNAs originate from repeats (57.3% to 66.7%), with the largest overlap observed for brain-specific ncRNAs. **b.** Distribution of repeats present at the origin of ncRNAs including major repeat families. The distribution is similar for the two groups, with LINE /L1 and LINE/L2 and SINE/Alu and SINE/MIR covering almost 70% of all the repeat families at origin. DNA/TcMar-Tigger elements are over-represented in brain-specific ncRNAs. **c.** Comparison of repeats present in the body of ncRNA transcripts. The percentage of brain-specific and non-tissue-specific ncRNAs containing at least one repeat within the transcript is shown. The vast majority of the ncRNAs overlap at least one repeat (85.1% to 91.4%), with brain-specific ncRNAs showing the largest overlap. **d.** Distribution of repeats present in the body of ncRNAs. The distribution of major repeat families is similar across the two groups and the most represented families are LINE/L1 and LINE/L2 and SINE/Alu and SINE/MIR. LINE/L1 repeats are depleted from non-tissue-specific ncRNAs, which are enriched for SINE/Alu repeats. Brain-specific ncRNAs are significantly depleted of SINE/Alu and SINE/MIR repeats. **e.** Percentage of ncRNA sequence derived from repeat elements for brain-specific and non-tissue-specific ncRNAs: for both groups about 50% of the sequence is repeat-derived. **f.** Distribution of repeat-derived sequence across the major repeat families investigated in this study. Although the pattern is similar to what obtained by counting the repeats in the gene body of the ncRNAs of the two groups (Figure 2d) the differences in terms of bp composition are not significant.

**Figure 2**



## Methods

From the published microarray dataset <sup>39</sup> we extracted all the ncRNAs expressed in a single tissue and in multiple tissues, as reported in the original publication. After evaluating the number of tissue-specific ncRNAs for each of the 12 tissues assessed in [39] we focused our analyses on the two sets of ncRNAs expressed exclusively in brain and in more than one tissue (referred to in the manuscript as “brain-specific” and “non-tissue-specific” respectively). To perform annotations “at origin” the 5' end of each tissue-specific ncRNA was extended upstream by 100 nucleotides. All gene annotations were based on GENCODE v17. Annotation with respect to repeats was performed using RepeatMasker (downloaded from UCSC genome browser - <http://genome.ucsc.edu/> - on 2014/01/30) and significance of differences between groups (ncRNAs overlapping/non-overlapping repeats and proportion of distinct repeat families at origin and in the gene body) were evaluated using Fisher exact test. The amount of repeat-derived sequence for brain- and non-tissue-specific ncRNAs was derived from ncRNA annotations provided in [39] and intersection files. Significance of differences in the distribution of bps covered by distinct repeat families per transcript were performed using Mann-Whitney U test. Intersection with ncHARs was done considering a 5kb-long window. All intersections were performed using closestBed, intersectBed or windowBed from bedTools suite <sup>56</sup>. Functional enrichment and gene up-regulation analyses were performed using DAVID <sup>57, 58</sup>. All plots and statistical tests were performed in R (<http://www.r-project.org/>).

## Acknowledgements

AS is supported by the National Institute for Health Research (NIHR) funded Biomedical Research Centre at Guy's and St. Thomas' Trust and Kings College London, United Kingdom. We thank S. Saxena from the City University London for careful reading of this manuscript.

## References

1. Carrieri, C. et al. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* 491, 454-7 (2012).
2. Khalil, A. M. et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* 106, 11667-72 (2009).
3. Saxena, A. & Carninci, P. Long non-coding RNA modifies chromatin: epigenetic silencing by long non-coding RNAs. *Bioessays* 33, 830-9 (2011).
4. Katayama, S. et al. Antisense transcription in the mammalian transcriptome. *Science* 309, 1564-6 (2005).
5. Kim, T. K. et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465, 182-7 (2010).
6. Orom, U. A. et al. Long noncoding RNAs with enhancer-like function in human cells. *Cell* 143, 46-58 (2010).
7. Washietl, S. et al. RNACode: robust discrimination of coding and noncoding regions in comparative sequence data. *Rna* 17, 578-94 (2011).
8. Chooniedass-Kothari, S. et al. The steroid receptor RNA activator is the first functional RNA encoding a protein. *FEBS Lett* 566, 43-7 (2004).
9. Dinger, M. E., Gascoigne, D. K. & Mattick, J. S. The evolution of RNAs with multiple functions. *Biochimie* 93, 2013-8 (2011).
10. Kondo, T. et al. Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis. *Science* 329, 336-9 (2010).
11. Candeias, M. M. et al. P53 mRNA controls p53 activity by managing Mdm2 functions. *Nat Cell Biol* 10, 1098-105 (2008).
12. Poliseno, L. et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 465, 1033-8 (2010).
13. Bernstein, B. E. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74 (2012).
14. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* 489, 109-13 (2012).
15. Djebali, S. et al. Landscape of transcription in human cells. *Nature* 489,



- 101-8 (2012).
16. Derrien, T. et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 22, 1775-89 (2012).
  17. Banfai, B. et al. Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res* 22, 1646-57 (2012).
  18. Mercer, T. R., Dinger, M. E., Sunken, S. M., Mehler, M. F. & Mattick, J. S. -specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A* 105, 716-21 (2008).
  19. Ponjavic, J., Oliver, P. L., Lunter, G. & Ponting, C. P. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet* 5, e1000617 (2009).
  20. Lipovich, L. et al. Developmental Changes in the Transcriptome of Human Cerebral Cortex Tissue: Long Noncoding RNA Transcripts. *Cereb Cortex* (2013).
  21. Mercer, T. R. et al. Long noncoding RNAs in neuronal-glial fate -specification and oligodendrocyte lineage maturation. *BMC Neurosci* 11, 14 (2010).
  22. Mercer, T. R. et al. Noncoding RNAs in Long-Term Memory Formation. *Neuroscientist* 14, 434-45 (2008).
  23. Bernard, D. et al. A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. *Embo J* 29, 3082-93 (2010).
  24. Lipovich, L. et al. Activity-dependent human brain coding/noncoding gene regulatory networks. *Genetics* 192, 1133-48 (2012).
  25. Modarresi, F. et al. Inhibition of natural antisense transcripts in vivo results in genespecific transcriptional upregulation. *Nat Biotechnol* 30, 453-9 (2012).
  26. Arisi, I. et al. Gene expression biomarkers in the brain of a mouse model for Alzheimer's disease: mining of microarray data by logic classification and feature selection. *J Alzheimers Dis* 24, 721-38 (2011).
  27. Chung, D. W., Rudnicki, D. D., Yu, L. & Margolis, R. L. A natural antisense

- transcript at the Huntington's disease repeat locus regulates HTT expression. *Hum Mol Genet* 20, 3467-77 (2011).
28. Mus, E., Hof, P. R. & Tiedge, H. Dendritic BC200 RNA in aging and in Alzheimer's disease. *Proc Natl Acad Sci U S A* 104, 10679-84 (2007).
  29. Johnson, R. et al. The Human Accelerated Region 1 noncoding RNA is repressed by REST in Huntington's disease. *Physiol Genomics* (2010).
  30. Faghihi, M. A. et al. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat Med* 14, 723-30 (2008).
  31. Saxena, A., Tang, D. & Carninci, P. piRNAs warrant investigation in Rett Syndrome: An omics perspective. *Dis Markers* 33, 261-75 (2012).
  32. Cabili, M. N. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and -specific subclasses. *Genes Dev* 25, 1915-27 (2011).
  33. Ponjavic, J., Ponting, C. P. & Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* 17, 556-65 (2007).
  34. Tsai, M. C. et al. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 329, 689-93 (2010).
  35. Rinn, J. L. et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129, 1311-23 (2007).
  36. Gupta, R. A. et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464, 1071-6 (2010).
  37. Nie, Y. et al. Long non-coding RNA HOTAIR is an independent prognostic marker for nasopharyngeal carcinoma progression and survival. *Cancer Sci* 104, 458-64 (2013).
  38. Schorderet, P. & Duboule, D. Structural and functional differences in the long non-coding RNA hotair in mouse and human. *PLoS Genet* 7, e1002071 (2011).
  39. Nielsen, M. M. et al. Identification of expressed and conserved human noncoding RNAs. *Rna* 20, 236-51 (2014).
  40. Franceschini, A. et al. STRING v9.1: protein-protein interaction networks,

- with increased coverage and integration. *Nucleic Acids Res* 41, D808-15 (2013).
41. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* 409, 860-921 (2001).
  42. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* 10, 691-703 (2009).
  43. Garcia-Perez, J. L. et al. Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. *Nature* 466, 769-73 (2010).
  44. Faulkner, G. J. et al. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* 41, 563-71 (2009).
  45. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol* 13, R107 (2012).
  46. Muotri, A. R. et al. L1 retrotransposition in neurons is modulated by MeCP2. *Nature* 468, 443-6 (2010).
  47. Baillie, J. K. et al. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* 479, 534-7 (2011).
  48. Evrony, G. D. et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* 151, 483-96 (2012).
  49. Cohen, C. J., Lock, W. M. & Mager, D. L. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* 448, 105-14 (2009).
  50. Conley, A. B., Piriyaopongsa, J. & Jordan, I. K. Retroviral promoters in the human genome. *Bioinformatics* 24, 1563-7 (2008).
  51. Bird, C. P. et al. Fast-evolving noncoding sequences in the human genome. *Genome Biol* 8, R118 (2007).
  52. Bush, E. C. & Lahn, B. T. A genome-wide screen for noncoding elements important in primate evolution. *BMC Evol Biol* 8, 17 (2008).
  53. Pollard, K. S. et al. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443, 167-72 (2006).
  54. Prabhakar, S., Noonan, J. P., Paabo, S. & Rubin, E. M. Accelerated

- evolution of conserved noncoding sequences in humans. *Science* 314, 786 (2006).
55. Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L. & Pollard, K. S. Many human accelerated regions are developmental enhancers. *Philos Trans R Soc Lond B Biol Sci* 368, 20130025 (2013).
  56. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841-2 (2010).
  57. Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4, 44-57 (2009).
  58. Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37, 1-13 (2009).
  59. Pang, K. C. et al. RNADB – a comprehensive mammalian noncoding RNA database. *Nucleic Acids Res* 33, D125-30 (2005).
  60. Luo, H. et al. Comprehensive characterization of 10,571 mouse large intergenic noncoding RNAs from whole transcriptome sequencing. *PLoS One* 8, e70835 (2013).
  61. Qureshi, I. A. & Mehler, M. F. Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. *Nat Rev Neurosci* 13, 528-41 (2012).
  62. Iyengar, B. R. et al. Non-coding RNA interact to regulate neuronal development and function. *Front Cell Neurosci* 8, 47 (2014).
  63. Ng, S. Y., Johnson, R. & Stanton, L. W. Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *Embo J* 31, 522-33 (2012).
  64. Ng, S. Y., Bogu, G. K., Soh, B. S. & Stanton, L. W. The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis. *Mol Cell* 51, 349-59 (2013).

# **Chapter 6**

## **General discussion**



## 6.1 Results summary

The broad aims of this thesis were to gain insight into brain-specific and region-specific transcriptional features and to create a high-resolution expression profiling atlas across distinct regions of the human aged CNS. To address these questions, we started out with a pilot project aiming at profiling TSSs in 5 regions of the CNS in human aged donors and investigate the relationships between methylation and expression. The results of this study are presented in Chapter 2. In summary, using CAGE as expression profiling technique, we found that a large proportion of the genes expressed show evidence of alternative promoter usage, which we hypothesized to be a major mechanism in establishing transcriptional diversity across distinct regions: 80% of the TSSs differentially expressed across regions were alternative transcription initiation sites for genes that also expressed a promoter similarly in all regions. We observed limited correlation between methylation and expression levels; in particular differential methylation explained a limited proportion of the differential expression observed across regions (only 5% of the expressed TCs showed differences in expression that could be explained by methylation effects). Additionally we unexpectedly found that 75% of the methylation signal was derived from gene bodies. Surprisingly the differentially expressed TFs showed enrichment in functional terms related to neural development, despite the fact that the study was performed in aged donors and that neurons are terminally differentiated post-mitotic cells. In Chapter 3, as part of the FANTOM5 consortium, we participated in the creation of a revolutionary CAGE-based promoterome atlas, providing expression profiles for 701 primary cells (573 from human and 128 from mouse) and 423 tissues (152 human post-mortem and 271 mouse developmental samples), complemented with 250 human cancer cell lines. Notably, with this comprehensive dataset we were able to provide evidence of expression for 91% of the human protein-coding genes. Interestingly only a limited number of them (6%) could be considered truly housekeeping, in the sense that they were expressed ubiquitously (in more than 50% of the samples) and uniformly (less than 10-fold difference between median and maximum

expression). A crucial result of the study is that cell-specific expression is achieved by a combination of proximal TF motifs and highly specific enhancers and this mechanism is shared by both CpG Island (CGI) and non-CGI promoters; additionally many CGI promoters (54%) and most TATA promoters (98%) had non-ubiquitous expression profiles. In Chapter 4 we moved the focus to the CNS samples available in the FANTOM 5 tissue collection, which includes frontal, temporal, occipital and parietal cortices, amygdala and hippocampus, caudate, putamen, thalamus, globus pallidus, locus coeruleus, substantia nigra, medulla oblongata, spinal cord and cerebellum and represent an expansion of the data presented in Chapter 2. Using these data we were able to show that tissues of the CNS have a distinctive expression signature with respect to all other tissues available in the collection and the defining elements of this specific signature are not only protein-coding genes and TFs, but also ncRNAs and intergenic peaks, that represent putative new transcripts. Additionally CNS tissues were characterized by higher transcriptional complexity, to which non-coding transcripts significantly contributed. We showed that TSSs up-regulated in brain are characterized by a specific transcriptional context, enriched in simple and low-complexity repeats and CG-rich regions. When investigating region-specific expression signatures we identified 4 major co-expression groups that associated functionally or anatomically related regions: e.g. striatum was one of the expression groups, composed of caudate and putamen; we observed similar expression signature across the cortex samples and amygdala and hippocampus, co-expression group that we denoted as “cortex-limbic system” group. Interestingly, analogous co-expression groups were also found when clustering the expression profiles of TFs and long non-coding RNAs (lncRNAs) up-regulated in the CNS tissues. Among transcripts with restricted expression patterns there were several ncRNAs and novel transcripts, which even though novel were supported by good evidence that they are genuine transcripts. Given this evidence that ncRNAs are abundantly transcribed in the tissues of the CNS and contribute to regionally biased expression, we decided to further study the characteristics of brain-specific ncRNAs from another perspective. In Chapter 5, using a previously published custom microarray dataset, we investigated the landscape of ncRNAs



expressed in 12 distinct human tissues and showed that brain expresses the largest number of tissue-specific ncRNA. We found that protein-coding genes neighboring brain-specific ncRNAs are enriched for genes up-regulated in brain, however they didn't show enrichment in GO terms associated to brain function. We then investigated the representation of repetitive elements at the origin and in the gene body of brain-specific ncRNAs and found that significantly more brain-specific ncRNAs originate and contain repeats with respect to non-tissue-specific ncRNAs. Investigation of the distribution across repeat families in the two groups showed that DNA/TcMar-Tigger elements are over-represented at the origin of brain-specific ncRNAs, while LINE/L1 repeats are enriched in their gene body, which is on the other hand depleted of SINE/Alu elements.

## **6.2 Discussion**

With the work presented in this thesis we aimed at contributing to the knowledge about transcription in brain in general, and in distinct regions of it in particular. When we started out in 2009, comprehensive expression studies of the human CNS were limitedly available, typically performed on poorly characterized brain samples (e.g. age of the donor unknown) or limited to a very specific context, in terms of type of samples or regions of the CNS assessed (a single region relevant in the pathogenesis of a specific neurodegenerative disease, such as e.g. substantia nigra in PD). Dedicated brain studies existed, typically performed with microarray technology (e.g. (Roth et al. 2006)) and therefore burdened by the limitation of being able to profile transcription only for the genes spotted on the array. For these reasons we set out with the long term project of creating a high-resolution expression atlas with the benefits of a novel technology and with well characterized samples, to be used in future work as controls for studies focusing on brain of patients affected by neurodegenerative diseases. In parallel to our work, in the last few years very interesting studies were performed to systematically profile transcription across distinct regions of the human CNS, at

different time-points across development, maturation and aging, which brought novel information and important resources to dissect CNS expression. In particular, Colantuoni et al. provided a detailed study of human prefrontal cortex expression dynamics and its relationship with genomic variation in a comprehensive series of post-mortem samples ranging in age from fetal to aged (Colantuoni et al. 2011). Kang et al. provided a detailed cross-region and cross-age expression profiling study based on exon microarrays, complemented with information on genomic variation (Kang et al. 2011). Finally Hawrylycz et al. and Miller et al. presented extremely detailed atlases assessing expression in a very large array of anatomical regions of the CNS, for both adult (Hawrylycz et al. 2012) and developmental (Miller et al. 2014) samples. These works, although exhaustive in terms of time-points and regions assessed, are still based on microarray platforms, which dramatically limit the possibility to acquire information about currently unknown transcripts. To overcome this limitation, we designed our expression profile studies based on a technique that is independent on currently known annotations, and is therefore well suited for the discovery of novel transcripts.

With our work we show that the CNS is characterized by specific global transcriptional features. Importantly tissues of the CNS are characterized by higher transcriptional complexity with respect to other tissues. There is probably no doubt that brain is the most complex organ of the human body. In the past years there have been several studies suggesting that brain has a relatively high number and large variety of genes expressed (Ramsköld et al. 2009), that it is characterized by high rates of alternative splicing (Yeo et al. 2004) and that brain-derived NGS libraries have high transcriptional complexity, typically surpassed or similar in magnitude to testis (Jongeneel et al. 2005; Ramsköld et al. 2009). Although we didn't find evidence that CNS expresses the largest number of genes (not shown), the results presented in Chapter 4 clearly support the concept that the transcriptional complexity of the CNS tissues exceeds that of the other tissues. One possible explanation for this is that CNS tissues are composed of a very heterogeneous mixture of cell types. It is however difficult to test this hypothesis, even in the largest context of the FANTOM5 promoterome atlas presented in

Chapter 3, because of the limited availability of distinct neuronal and glial primary cells in the collection.

Brain tissues were also suggested in past studies to be characterized by a specific “transcriptional context”, i.e. transcription for brain-specific genes initiates in genomic regions characterized by enrichment in CG rich regions and specific repeat elements (Faulkner et al. 2009; Roider et al. 2009; Xu et al. 2010). In line with these observations, in Chapter 4 we show that TSSs up-regulated in the CNS often originate in CG-rich regions as well as in simple and low complexity repeats. That a relationship between brain-specific transcription and repeats distribution exists, is further supported by the study presented in Chapter 5, where we show that brain-specific ncRNAs generally originate and contain repeats more frequently than non-tissue-specific ncRNAs and that specific repeats classes seem to be involved. This is particularly interesting in light of the fact that expressed repeats have been shown to be non-randomly distributed across the genome (Xu et al. 2010; Kelley and Rinn 2012) and are suggested to be actively involved in the regulation of gene expression, by providing alternative promoters (Cohen et al. 2009), alternative exons (Shen et al. 2011) or by being associated to specific transcript classes and tissue-specific expression patterns (Kelley and Rinn 2012). The results presented in Chapters 3 and 4 clearly show that the CNS system has an overall transcriptional profile that neatly separates it from other tissues, as suggested in the first large expression profiling work performed on a comprehensive selection of different regions of the human post-mortem CNS and other tissues (Roth et al. 2006). Interestingly this is true not only at the level of protein-coding genes, which would be somehow expected since very specific components are needed and produced in the CNS (e.g. all the molecular components that participate in the synthesis and secretion of neurotransmitters). Strikingly, instead, the CNS samples are set apart from all other tissues in the human body considering also expression from less well established transcript classes, such e.g. lncRNAs and CAGE-defined intergenic peaks, most likely indicating novel transcripts or TSSs. Among the protein-coding genes, TFs in particular are clearly determining CNS signature, in line with the observation presented in Chapter 3 that related primary cells clearly cluster together based on

TF expression. Unexpectedly, part of the TFs up-regulated in the CNS identified in Chapter 4 are not pointed out by literature as key elements for the CNS function, while the remaining are well known for their crucial role in brain development. In terms of TFs, in Chapter 2 we find enrichment of neurodevelopmental functional terms in the set of TFs differentially expressed across regions: this might be surprising at first sight, but it is reasonable that TFs that are important for neural development keep playing a role in the aged brain, although it would be interesting to investigate their exact functional role there.

In Chapter 2 we also discovered that abundant methylation signal in the distinct brain regions profiled is located in gene bodies and that there is limited correlation between methylation and expression. In retrospective, several reports are now available showing how the relationship between methylation and expression is more intricate than originally established (reviewed in (Jones 2012)). In particular it is suggested that methylation in the gene bodies is a widespread phenomenon linked to the regulation of alternative promoter usage and alternative exon inclusion (Maunakea et al. 2010; Maunakea et al. 2013).

Moving towards a more global picture of transcription, we participated in the creation of the comprehensive promoterome atlas presented in Chapter 3. This work provides the scientific community with a powerful resource, thanks to the creation of several easily accessible integrative analyses. User-friendly tools allow for the easy access to relevant TFs and putative regulators in specific cell types and tissues, which can be readily used to formulate hypothesis to be tested at the bench. Additionally a guilt-by-association approach provides putative functional annotation to transcripts for which current information is limited. An important observation that emerges from the study is that only a very limited number of genes can be considered as truly housekeeping, which raises important reflections in the view that housekeeping genes are often used to perform internal normalizations. It is also important to note that cancer cell lines expression profiles tend to be divergent from their tissues of origin, while primary cell expression profiles cluster together with their tissue counterpart. This suggests that primary cells should be the system of choice when studying phenotypes that are relevant for a specific tissue. Given the improvement in induced pluripotent stem cell

(iPSC) technology and the advent of standardized characterization guidelines for differentiated cells, a viable alternative that is starting to emerge is given also by differentiated iPSC.

The second aim of this thesis was to create a high resolution expression profile atlas of distinct brain regions from aged donors. This aim was first approached in Chapter 2 and was then expanded in Chapter 4, with the creation of a high-resolution expression atlas for 15 regions of the human CNS. We identified extensive differences in expression across regions: unexpectedly we didn't observe expression signatures specific for each of the regions profiled but rather transcription profiles that tend to be similar between developmentally, functionally or morphologically related regions, with cerebellum showing the most distinctive expression signature and the largest number of differentially expressed TCs, consistently with results presented in (Kang et al. 2011; Hawrylycz et al. 2012). One of the most innovative contributions of the works presented here is the identification of large sets of brain-specific and region-specific ncRNAs and novel transcripts expressed. In the last few years awareness that several ncRNAs are key actors in the largest variety of biological processes started to emerge, such as regulation of gene expression through chromatin modifications (Khalil et al. 2009; Saxena and Carninci 2011) and enhancer activity through transcribed RNAs (Kim et al. 2010) and lncRNAs (Ørom et al. 2010). Additionally it has been recently shown in mouse that long intergenic ncRNAs are required for life and brain development (Sauvageau et al. 2013). From the studies presented in Chapters 2, 4 and 5 it becomes apparent that brain hosts expression for a large variety of ncRNAs and that these have expression patterns comparable to that of protein-coding genes. Additionally the study presented in Chapter 5 suggests that they have specific characteristics in terms of repeat expression. Certain types of repeat elements were proposed to have specific active roles in the regulation of gene expression (Cohen et al. 2009; Xu et al. 2010; Shen et al. 2011; Kelley and Rinn 2012). A particularly intriguing example of transcriptional regulation mediated by a repeat element is given by (Carrieri et al. 2012). This study shows that the expression of the gene *Uchl1*, which human homologue is a susceptibility gene for PD (Maraganore et al. 2004), is regulated post-transcriptionally by an antisense

non-coding transcript and that the regulation is mediated by an inverted SINE repeat embedded in the ncRNA. This suggests an intriguing mechanism by which ncRNAs might be involved in high-level regulation of expression that might be relevant for neurodegeneration.

### 6.3 Limitations

There are several aspects in the works presented in this thesis that should be considered with critical eye. To start with, we mainly focused on expression profiling based on CAGE technique. Although unmatched to precisely identify the genomic location of TSSs and therefore ideal to characterize alternative TSSs differentially used in distinct tissues or brain regions, it is affected by the inherent limitation of being completely focused on transcription initiation events and not allowing for the investigation of the full length nature of the transcripts identified, such as e.g. intron-exon alternative structures. RNA-seq (Mortazavi et al. 2008) is an alternative genome-wide expression profiling technique widely used in the production phase of the ENCODE project. This technique doesn't allow for the same precision in the identification of TSSs, but provides information about the full length structure of the transcripts, although building transcripts *de novo* is still a challenging problem, specially for samples with complex splicing patterns and high sequencing depth (Steijger et al. 2013). It has been recently shown that expression levels assessed by CAGE and RNA-seq are largely comparable, and therefore the two techniques can be efficiently used in combination to improve information about new or incomplete gene models (Kawaji et al. 2014). Considering the technology itself, it is important to notice that CAGE is not suitable for the profiling of short RNAs without a specific adaptation of the protocol, since typical cDNA preparation protocols exclude RNAs shorter than 500bp (Wu et al. 2013).

It is important to highlight that these RNA-based studies provide indeed information at the RNA level, but this doesn't always directly correlate with protein

expression levels (Gry et al. 2009). Although recent technological advances, the large-scale identification of proteins in a sample is still a challenging problem and typically provides information for a limited number of peptides. A viable alternative is ribosome profiling (Ingloia et al. 2009), which uses RNA sequencing technology to monitor protein synthesis. Although this is clearly an important step forward towards genome-wide protein profiling, the technique cannot incorporate information about protein stability and degradation rates, and therefore still provides an incomplete picture.

Another point to be considered is the use in the studies presented here of post-mortem material. It's been suggested that expression profiles derived from post-mortem material can be affected by biases related to e.g. the agonal state (Li et al. 2004). In our studies, besides the observation that libraries prepared with material with low RNA integrity number (RIN) typically fail at the sequencing stage, we didn't observe systematic biases in expression directly correlated to post-mortem delay, RIN number or pH, all measures that can be altered due to agonal state. It is still possible, however, that biases in the data exist that we didn't account for yet. It is certainly true, for example, that the relative degradation rates across distinct RNAs influence RNA expression profile measures; however it was recently suggested that only heavy degradation has effects that can be measured by RNA-seq (Gallego Romero et al. 2014). It would be interesting to compare post-mortem brain material with surgically removed tissue: however on the one hand surgically removed tissue is difficult to access and is usually available in limited amounts that represent a technical challenge for an expression profiling study, on the other hand a tissue derived with such a procedure would still be subject itself to a specific stress, so it could in principle be affected by other but not less important biases. From this point of view, it is likely that an efficient workaround doesn't exist, and the only sensible approach is to carefully evaluate the presence of systematic biases in the data, and remove the samples from the analyses if such biases are identified and cannot be corrected at the analysis level.

In Chapter 4, although a large number of CNS regions were profiled, we were unable to identify expression signatures for each of them. This is partly consistent with other publications, reporting that the largest differences between CNS regions

are observed during prenatal development (Kang et al. 2012). There are however two additional possible explanations for this, besides the one suggesting that there is indeed no difference. It is possible that differences in expression are very small, and therefore we would need to expand the sample size in order to have a higher resolution. Alternatively, it is possible that all the differences between e.g. frontal and temporal cortex are limited to subsets of cells in the tissues profiled, and since we are measuring only averaged expression we are unable to detect them. This last possibility could be explored by undertaking laser capture or cell sorting approaches.

## **6.4 Future directions**

Importantly, the third aim of this thesis was the complementation of the region-specific expression atlas with data derived from patients affected by neurodegenerative diseases. We are currently in the process of collecting and analyzing expression data derived from different CNS regions of patients affected by familial FTD with underlying mutations in distinct genes. The plan is to analyze expression in a network framework, in order to identify pathways that are commonly and differently dysregulated across both distinct mutations and distinct regions. Besides dealing with already established pathways and interacting elements, taking advantage of correlation structures present in the data and using network analysis approaches (such as weighted gene co-expression network analysis – WGCNA (Fuller et al. 2007)) it is possible to identify novel interactors. On the one hand we aim at understanding the biological processes that through distinct mutations give rise to what can be broadly considered as the same phenotype (FTD). On the other hand by including in the study regions of the CNS that are affected at different grades of severity by the disease gives an important opportunity to gain insight into region-specific vulnerability to neurodegeneration. While expression profiling and network analysis identify potential pathway elements, appropriate systems are needed for validation and causal-relationship assignment. A powerful approach to investigate this is given by high-throughput



high-content cellular screens, that can be used to impose direct regulatory interactions in the elements of the pathway (Jain and Heutink 2010). In particular, experiments can be performed in iPSCs derived from patients and controls, which gives a readout in a cell-type specific system that is particularly close to the disease.

As a more global outlook, all the work presented in this thesis focused on transcription profiling. However this is only one aspect of the global picture and it is becoming clear that more comprehensive studies are needed to explain complex biological phenomena, such as transcription. Current lines of research are starting to migrate from transcriptome alone to other “-omes”, such as interactome, epigenome, etc. It is undoubtedly true that these novel layers of regulation that we are starting to dissect and understand play an important role in the CNS. For example, recent studies suggest that aberrant histone acetylation is the cause of abnormal transcriptional profiles observed in aging brain (Pirooznia and Elefant 2013). I find particularly intriguing the technique of chromosome conformation capture, which allows for the identification of the portions of the genome that physically interact in the three-dimensional space of the nucleus. Comprehensive studies in subsets of cell lines from the ENCODE project suggest that an important part of transcriptional regulation happens at the “neighborhood level”: e.g. enhancer sequences are physically close to their target genes, and when the physical interaction is disrupted, the enhancer activity is disrupted as well.

I also believe that an important challenge for future research is “integration”, not only at the level of datasets but also at the level of knowledge, using systems biology approaches that allow for the integration of datasets and information. Additionally, there is an incredible need to improve functional annotation. In the studies presented here we identified thousands of transcripts that are expressed specifically in brain or in specific super-groups of regions of the CNS; however, the information available about their function is very limited and often absent, which makes very complicated the formulation of hypothesis based on the data.

## 6.5 Conclusions

The aims of this thesis were:

1. to gain general insight into the dynamics of transcription in the CNS
2. to specifically create a high resolution expression profile atlas encompassing distinct regions of the CNS in aged donors
3. to extend the atlas to matched disease samples, in order to identify and functionally validate networks of co-expressed transcripts perturbed in disease

In Chapters 3, 4 and 5 we investigated transcriptional features and expression profiles of the CNS with respect to other tissues of the human body. We identified transcriptional contexts, transcriptional characteristics, TFs and ncRNAs that are CNS-specific, fulfilling the first aim of this thesis. In Chapter 2 we created a high-resolution expression atlas encompassing 5 regions of the aged human CNS, which we expanded in Chapter 4 by adding 10 additional anatomical regions. We identified several coding and non-coding transcripts that are expressed in a regionally biased manner, fulfilling the second aim of this thesis. The third aim is currently ongoing, and it is discussed as a future direction. Overall, with the work presented here, we provide high-resolution expression information about coding and non-coding genes expressed in the CNS and its distinct regions. We supply powerful resources for hypothesis making and testing and for gaining insight into mechanisms involved in the regulation of gene expression in brain. Additional work and data production is under way and will importantly complement the results already achieved. In particular the growing amount of high quality expression data available, along with new analysis and validation techniques, is importantly contributing to the dissection of CNS expression features and will improve our understanding of regional vulnerability in neurodegenerative diseases. Thanks to system biology perspectives and approaches, the next years will see the advent of new understanding in the pathogenesis of these diseases, opening the door to new therapeutic and preventive strategies.

## Acknowledgments

When I started the PhD I hadn't a very clear idea of what I was getting into. It's been a long path, which brought me to many places around the world, saw me changing into who I am today and allowed me to meet many fantastic people along the way, without whom I wouldn't be writing this thesis now.

In particular, many thanks to Peter Heutink, my supervisor. We've been working together for quite a while now, and you definitely had a big part in shaping my scientific thinking. I thank you a lot for giving me great opportunities, for teaching me how to be independent in my scientific thinking and working, for being patient in the moments of crisis and not getting too annoyed with reports/drafts/presentations/theses to read on short notice. You also showed me how to think BIG, and this is definitely an opportunity that doesn't happen every day.

In time I had a lot of good colleagues, that eventually became friends. Special thanks to Cornelis: you helped me with writing, you helped me with thinking and you also helped me a lot with complaining! I hope we can keep going like this for a long time! And thanks to Sasja, I hope you'll be at this stage soon as well, and thank you Iris, and Ashu, for the discussions and the dinners and specially for helping me taking decisions and move on when I get stuck in a spiral of thinking. Every time I don't know how to react you are the first persons I ask to: thanks for showing the way. And thank you Patrizia, a lot, for making me think straight when I start to freak out.

Many thanks are due to my family: Ma, Dudu, Nic and my super grandma, None Cila. You all have been supportive in my whole life in a very special way. You've given me freedom and trust, and I definitely wouldn't be here and wouldn't be who I am without you. You've been supportive in happy and unhappy times, encouraging me to do what must be done and finish what I started. And we discussed, we cooked, we walked, we biked, we hiked and we did all these nice things together: I quite regret the fact that we live so far apart now and my visits

are just short and busy. I guess I'd like to thank the smaller members of the family too, but it looks strange so I skip.

Nothing of this would have been possible without the GABBA Program, a very special doctoral program that I discovered by chance when I met Patricia in Trieste (thanks Pat for sending me here: it was a great thing to do!). In particular thanks to my co-supervisor Alexandre do Carmo and to Catarina, that besides being the coordinator of many aspects of GABBA-flow also became a very good friend. And, of course, thanks to GABBA12: we spent part of our lives together, and I didn't really imagine we could get so close in such a short time. Special thanks to Bebs: without you this thesis wouldn't be finished and it wouldn't be printed, or probably just printed wrong. Without even mentioning all the support you gave me along the way! And Cris, Di: meeting you was a gift and I hope we'll be able to keep catching up time to time, as we did so far!

One of the most unexpected places I ended up being is Japan: I never thought it's a place where people actually go. The time I spent there was very constructive and many people I met will always have a special place in my thoughts. In particular thanks to Piero, Alistair, Michiel and Carsten for being in a way or another supervising me for the time I spent there. You taught me a lot, and I thank you for this. Thanks to Thierry, because of all the discussions we had about science in general and life in particular. And very very special thanks to The One! Mo, meeting you was a great thing. We've been discussing, arguing, talking, hanging around and chatting chatting chatting. We shared bad moments and good ones, and we solved problems and had ideas and also finished something here and there. This PhD would have been quite a different experience if I didn't meet you, possibly it wouldn't have turned into a PhD at all. I am really happy we've kept being in touch for so many years, and I really hope this will stay the same even when our working paths will eventually diverge.

Puppi, Guido, Giada, Fonta, Chiara, and all other friends from Italy, with whom I have the chance to discuss a lot more than I would have thought few years back, but still less than I would like to. I grew up with you into my adult me, and I learned from you a lot. Your knowledge and view of the world and science is unique: though I traveled quite a bit by now, nowhere else I found somebody else as sharp

and interested in the strangest aspects of life and everything. I consider you my extended family, and you'll always have a special place in my hearth. One particular thank you to Guido: although our research paths ended up being quite different, I kept bugging you with questions about comparisons, statistics, Latex, presentations, posters and of, course, thesis: thank you for taking the time to think about my issues!

Last, but in truth first, very special thanks to Peter (my Peter): you've been thrown into this crazy life without knowing what you were heading for, and showed many times a patience and an understanding that are just incredible. For this I thank you, I thank you a lot. Every day I am happy to have you, and every day I spend somewhere else (and they are many, I know, sorry) I miss you. Te ljubim ogromno!

I thank all of you, also the ones that I didn't mention: in a way or another you contributed to this.



## References

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56-65.

Arrasate M, Finkbeiner S. 2012. Protein aggregates in Huntington's disease. *Exp Neurol* 238(1):1-11.

Arvey A, Agius P, Noble WS, Leslie C. 2012. Sequence and chromatin determinants of cell- type-specific transcription factor binding. *Genome Res* 22(9):1723-34.

Ashrafi G, Schwarz TL. 2013. The pathways of mitophagy for quality control and clearance of mitochondria. *Cell Death Differ* 20(1):31-42.

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, et al. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* 479(7374):534-7.

Bigio EH. 2013. Making the diagnosis of frontotemporal lobar degeneration. *Arch Pathol Lab Med* 137(3):314-25.

Brown TA. *Genomes*. 2nd edition. Oxford: Wiley-Liss; 2002. Chapter 1, The Human Genome.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. 2005. The transcriptional landscape of the mammalian genome. *Science* 309(5740):1559-63.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engström PG, Frith MC, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 38(6):626-35.

Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, Pesce E, Ferrer

I, Collavin L, Santoro C, et al. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* 491(7424):454-7.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437(7055):69-87.

Cohen CJ, Lock WM, Mager DL. 2009. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* 448(2):105-14.

Cunningham C. 2013. Microglia and neurodegeneration: the role of systemic inflammation. *Glia* 61(1):71-90.

Dauer W, Przedborski S. 2002. Parkinson's disease: mechanisms and models. *Neuron* 39(6):889-909.

Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, Green RD, Dekker J. 2006. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 16(10):1299-309.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* 489(7414):101-8.

ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447(7146):799-816.

ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57-74.

FANTOM Consortium, Suzuki H, Forrest AR, van Nimwegen E, Daub CO, Balwierz PJ, Irvine KM, Lassmann T, Ravasi T, Hasegawa Y, et al. 2009. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet* 41(5):553-62.



- Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* 41(5):563-71.
- Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Res.* 42(Database issue):D749-55.
- Fuller TF, Ghazalpour A, Aten JE, Drake TA, Lusis AJ, Horvath S. 2007. Weighted gene coexpression network analysis strategies applied to mouse weight. *Mamm Genome* 18(6-7):463-72.
- Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, et al. 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 462(7269):58-64.
- Gallego Romero I, Pai AA, Tung J, Gilad Y. 2014. RNA-seq: Impact of RNA degradation on transcript quantification. *BMC Biol* 12(1):42.
- Gamblin TC, Chen F, Zambrano A, Abraha A, Lagalwar S, Guillozet AL, Lu M, Fu Y, Garcia-Sierra F, LaPointe N, et al. 2003. Caspase cleavage of tau: linking amyloid and neurofibrillary tangles in Alzheimer's disease. *Proc Natl Acad Sci U S A.* 100(17):10032-7.
- Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, et al. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature* 489(7414):91-100.
- Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428(6982):493-521.
- Glass CK, Saijo K, Winner B, Marchetto MC, Gage FH. 2010. Mechanisms underlying inflammation in neurodegeneration. *Cell* 140(6):918-34.
- Glenner GG, Wong CW. 1984. Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein. *Biochem Biophys Res Commun* 120(3):885-90.
- Goate A, Chartier-Harlin MC, Mullan M, Brown J, Crawford F, Fidani L, Giuffra L,

Haynes A, Irving N, James L, et al. 1991. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* 349(6311):704-6.

Gry M, Rimini R, Strömberg S, Asplund A, Pontén F, Uhlén M, Nilsson P. 2009. Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics* 10:365.

Haass C, Selkoe DJ. 2007. Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid beta-peptide. *Nat Rev Mol Cell Biol* 8(2):101-12.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 22(9):1760-74.

Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, van de Lagemaat LN, Smith KA, Ebbert A, Riley ZL, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* 489(7416):391-9.

Hetz C and Mollereau B. 2014. Disturbance of endoplasmic reticulum proteostasis in neurodegenerative diseases. *Nat Rev Neurosci.* 15(4):233-49.

Hutton M, Lendon CL, Rizzu P, Baker M, Froelich S, Houlden H, Pickering-Brown S, Chakraverty S, Isaacs A, Grover A, et al. 1998. Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature* 393(6686):702-5.

Hyman BT, Van Hoesen GW, Damasio AR, Barnes CL. 1984. Alzheimer's disease: cell-specific pathology isolates the hippocampal formation. *Science* 225(4667):1168-70.

Ingolia NT, Ghaemmamghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324(5924):218-23.

International HapMap Consortium. 2003. The International HapMap Project. *Nature* 426(6968):789-96.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431(7011):931-45.

Jackson WS. 2014. Selective vulnerability to neurodegenerative disease: the curious case of Prion Protein. *Dis Model Mech* 7(1):21-9.

Johri A and Beal MF. 2012. Mitochondrial dysfunction in neurodegenerative diseases. *J Pharmacol Exp Ther* 342(3):619-30.

Jongeneel CV, Delorenzi M, Iseli C, Zhou D, Haudenschield CD, Khrebtukova I, Kuznetsov D, Stevenson BJ, Strausberg RL, Simpson AJ, Vasicek TJ. 2005. An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res* 15(7):1007-14.

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AM, Pletikos M, Meyer KA, Sedmak G, et al. 2011. Spatio-temporal transcriptome of the human brain. *Nature* 478(7370):483-9.

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, et al. 2005. Antisense transcription in the mammalian transcriptome. *Science* 309(5740):1564-6.

Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, et al. 2001. Functional annotation of a full-length mouse cDNA collection. *Nature* 409(6821):685-90.

Kawaji H, Lizio M, Itoh M, Kanamori-Katayama M, Kaiho A, Nishiyori-Sueki H, Shin JW, Kojima-Ishiyama M, Kawano M, Murata M, et al. 2014. Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res* 24(4):708-17.

Kelley D, Rinn J. 2012. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol* 13(11):R107.

Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, Regev A, Lander ES, Rinn JL. 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* 106(28):11667-72.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465(7295):182-7.

Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, et al. 2006. CAGE: cap analysis of gene expression. *Nat Methods* 3(3):211-22.

Lagier-Tourenne C, Polymenidou M, Cleveland DW. 2010. TDP-43 and FUS/TLS: emerging roles in RNA processing and neurodegeneration. *Hum Mol Genet* 19(R1):R46-64.

Lagier-Tourenne C, Baughn M, Rigo F, Sun S, Liu P, Li HR, Jiang J, Watt AT, Chun S, Katz M, et al. 2013. Targeted degradation of sense and antisense C9orf72 RNA foci as therapy for ALS and frontotemporal degeneration. *Proc Natl Acad Sci U S A* 110(47):E4530-9.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822):860-921.

Lander ES. 2011. Initial impact of the sequencing of the human genome. *Nature* 470(7333):187-97.

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, et al. 2012. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148(1-2):84-98.

Li JZ, Vawter MP, Walsh DM, Tomita H, Evans SJ, Choudary PV, Lopez JF, Avelar A, Shokoohi V, Chung T, et al. 2004. Systematic changes in gene expression in post mortem human brains associated with tissue pH and terminal medical conditions. *Hum Mol Genet* 13(6):609-16.

Lin MT, Beal MF. 2006. Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases. *Nature* 443(7113):787-95.

Maraganore DM, Lesnick TG, Elbaz A, Chartier-Harlin MC, Gasser T, Krüger R, Hattori N, Mellick GD, Quattrone A, Satoh J, et al. 2004. UCHL1 is a Parkinson's disease susceptibility gene. *Ann Neurol* 55(4):512-21.

- Mattson MP and Magnus T. 2006. Ageing and neuronal vulnerability. *Nat Rev Neurosci* Apr;7(4):278-94.
- Matus S, Glimcher LH, Hetz C. 2011. Protein folding stress in neurodegenerative diseases: a glimpse into the ER. *Curr Opin Cell Biol* 23(2):239-52.
- Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337(6099):1190-5.
- Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS. 2008. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A*. 105(2):716-21.
- Millecamps S, Julien JP. 2013. Axonal transport deficits and neurodegenerative diseases. *Nat Rev Neurosci*. 14(3):161-76.
- Miller JA, Ding SL, Sunkin SM, Smith KA, Ng L, Szafer A, Ebbert A, Riley ZL, Royall JJ, Aiona K et al. 2014. Transcriptional landscape of the prenatal human brain. *Nature* 508(7495):199-206.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7):621-8.
- Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520-62.
- Naidoo N, Pawitan Y, Soong R, Cooper DN, Ku CS. 2011. Human genetics and genomics a decade after the release of the draft sequence of the human genome. *Hum Genomics* 5(6):577-622.
- Nixon RA. 2013. The role of autophagy in neurodegenerative disease. *Nat Med*. 19(8):983- 97.
- Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420(6915):563-73.

Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zyt-nicki M, Notredame C, Huang Q, Guigo R, Shiekhata R. 2010. Long noncoding RNAs with enhancer-like function in human cells. *Cell* 143(1):46-58.

Palikaras K, Tavernarakis N. 2012. Mitophagy in neurodegeneration and aging. *Front Genet* 3:297.

Pirooznia SK and Elefant F. 2013. Targeting specific HATs for neurodegenerative disease treatment: translating basic biology to therapeutic possibilities. *Front Cell Neurosci* 7:30.

Polymeropoulos MH, Lavedan C, Leroy E, Ide SE, Dehejia A, Dutra A, Pike B, Root H, Rubenstein J, Boyer R, et al. 1997. Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* 276(5321):2045-7.

Przedborski S, Vila M, Jackson-Lewis V. 2003. Neurodegeneration: what is it and where are we? *J Clin Invest*. 111(1):3-10.

Ramsköld D, Wang ET, Burge CB and Sandberg R. 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* 5(12):e1000598.

Robberecht W, Philips T. 2013. The changing scene of amyotrophic lateral sclerosis. *Nat Rev Neurosci* 14(4):248-64.

Roider HG, Lenhard B, Kanhere A, Haas SA, Vingron M. 2009. CpG-depleted promoters harbor tissue-specific transcription factor binding signals--implications for motif over representation analyses. *Nucleic Acids Res* 37(19):6305-15.

Ross CA, Tabrizi SJ. 2011. Huntington's disease: from molecular pathogenesis to clinical treatment. *Lancet Neurol* 10(1):83-98.

Roth RB, Hevezi P, Lee J, Willhite D, Lechner SM, Foster AC, Zlotnik A. 2006. Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics* 7(2):67-80.

Sanyal A, Lajoie BR, Jain G, Dekker J. 2012. The long-range interaction landscape of gene promoters. *Nature* 489(7414):109-13.

Sauvageau M, Goff LA, Lodato S, Bonev B, Groff AF, Gerhardinger C, Sanchez-Gomez DB, Hacisuleyman E, Li E, Spence M, et al. 2013. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* 2:e01749.

Saxena A, Carninci P. 2011. Long non-coding RNA modifies chromatin: epigenetic silencing by long non-coding RNAs. *Bioessays* 33(11):830-9.

Saxena S, Caroni P. 2011. Selective neuronal vulnerability in neurodegenerative diseases: from stressor thresholds to degeneration. *Neuron* 71(1):35-48.

Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. 2012. Linking disease associations with regulatory information in the human genome. *Genome Res* 22(9):1748-59.

Shen S, Lin L, Cai JJ, Jiang P, Kenkel EJ, Stroik MR, Sato S, Davidson BL, Xing Y. 2011. Widespread establishment and regulatory impact of Alu exons in human genes. *Proc Natl Acad Sci U S A* 108(7):2837-42.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29(1):308-11.

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, et al. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A*. 100(26):15776-81.

Jain S, Heutink P. 2010. From single genes to gene networks: high-throughput-high-content screening for neurological disease. *Neuron* 68(2):207-17.

Soto C. 2003. Unfolding the role of protein misfolding in neurodegenerative diseases. *Nat Rev Neurosci* 4(1):49-60.

Stamatoyannopoulos JA. 2012. What does our genome encode? *Genome Res* 22(9): 1602- 1611.

Steijger T, Abril JF, Engström PG, Kokocinski F; RGASP Consortium, Abril JF, Akerman M, Alioto T, Ambrosini G, Antonarakis SE, et al. 2013. Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* 10(12):1177-84.

Sulzer D, Surmeier DJ. 2013. Neuronal vulnerability, pathogenesis, and Parkinson's disease. *Mov Disord* 28(6):715-24.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* 489(7414):75-82.

Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, et al. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* 22(9):1798-812.

Washietl S, Pedersen JS, Korbelt JO, Stocsits C, Gruber AR, Hackermüller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A, et al. 2007. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res* 17(6):852-64.

Wu J. *Post-Transcriptional Gene Regulation: RNA Processing in Eukaryotes*. John Wiley & Sons 2013.

Xu AG, He L, Li Z, Xu Y, Li M, Fu X, Yan Z, Yuan Y, Menzel C, Li N, et al. 2010. Intergenic and repeat transcription in human, chimpanzee and macaque brains measured by RNA- Seq. *PLoS Comput Biol* 6:e1000843.

Yeo G, Holste D, Kreiman G, Burge CB. 2004. Variation in alternative splicing across human tissues. *Genome Biol* 5(10):R74.